LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-865826

The Delicate Balance Redux: The Role of Nuclear Forces, Damage Limitation and Uncertainty in Future U.S.-China Crises

B. W. Bahney, B. Soper

June 20, 2024

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The Delicate Balance Redux: The Role of Nuclear Forces, Damage Limitation and Uncertainty in Future U.S.-China Crises

Benjamin Bahney[1] and Dr. Braden W. Soper
Lawrence Livermore National Laboratory

ABSTRACT

What is the impact of damage limitation capabilities like counterforce and missile defenses on deterrence, when their efficacy in stopping an adversary nuclear attack is uncertain? This is a key unanswered question to understand "how much is enough" for the United States to deter China and Russia in future nuclear crises. In this paper we extend an established, single move game theory model to capture the dynamics of two players in a nuclear crisis having varying damage limitation capabilities with uncertain effectiveness. Our model formalizes the logic of the "delicate balance" school of deterrence, which states that leverage in a crisis is driven by the risk each player can take with their combined strategic forces, and that those risks carry uncertainty as nuclear forces are hard to deliver against technologically advanced adversaries. Our model shows that damage limitation capabilities—even those with significant uncertainty around them like cyber or electronic warfare—can drive bargaining outcomes in an array of nuclear crises. We then apply these bargaining outcomes to the expected U.S.-China strategic balance as China builds out its nuclear force through 2035. We apply published force exchange models to determine the expected damage each side will be able to deliver, and we use these values to determine the likelihood that the U.S. can prevail in crises of varying stakes. Last, we show that U.S. policymakers have an array of options to improve future bargaining outcomes, evaluating how additional nuclear forces trade against improvements in damage limitation.

.

[1] Bahney1@llnl.gov

INTRODUCTION

How does crisis bargaining work between nuclear armed states when they can target each other's arsenals? This is a central question in the ongoing debate over U.S. strategic weapons decisions, given the expected growth of the Chinese arsenal to near parity with the U.S. by 2035, the fraying of formal arms control limits between the U.S. and Russia, and the possibility that China could shift to target military forces with its nuclear weapons. A recent assessment by the U.S. Congressional Strategic Posture Commission concluded that the U.S. military needs to increase the size of its nuclear arsenal to cope with these problems.[2] However, two separate groups of analysts have highlighted the high cost of expanding U.S. nuclear forces, and risk of that growth creating a competitive arms race spiral. These authors suggest, to differing extents, to shift the U.S. nuclear posture to target infrastructure and cities, rather than solely targeting adversary weapons via counterforce as the U.S. presently does.[3] But at the core of this debate is the unanswered question of how much the ability to limit the damage from attack by the other side—and the uncertainty that comes with it—factors into the deterrence equation in a crisis. If the bargaining value of damage limitation is low, policymakers should prioritize either nuclear force expansion or counter-value targeting. If the value of damage limitation is high, policymakers have additional options to shift the balance towards the U.S.' favor.

The debate over damage limitation and deterrence began in the early Cold War. Thomas Schelling famously likened nuclear crisis bargaining to a game of chicken where the balance of resolve dictates who bids higher, as both sides can theoretically annihilate one another due to conditions of nuclear stalemate.[4] But Albert Wohlstetter countered that stalemate is easily rolled

---

[2] Congressional Commission on the Strategic Posture of the United States 2023.
[3] Glaser, Acton, Fetter 2023; Lieber and Press 2023.
[4] Schelling 1960, p. 187.

back because second strike forces are extremely hard to safeguard and deliver, especially when they can be targeted by a damage-limiting first strike by the other side. This results in a "delicate nuclear balance" where the players are constantly concerned about the viability of their nuclear forces, and the balance of power is hard to determine due to secrecy and uncertainty.[5] In this case, competing nuclear armed states are sensitive to slight changes in both nuclear weapons and damage limitation postures, whereby damage limitation capabilities like missile defenses and counterforce can also create leverage in a crisis.[6] Herman Kahn elaborated on this idea to argue that if one side had a large and usable advantage over the other, it could threaten much more persuasively than its opponent to escalate or erupt to large scale war if a situation became desperate. The effect of the disparity in the strategic balance could thereby alter bargaining even at lower levels of escalation.[7] Both Wohlstetter and Kahn noted that the key factors in this balance are uncertain by nature, veiled by secrecy, and hard to measure.

These debates had an impact on policy. Recent scholarship has shown that in the Cold War, the U.S. pursued qualitative advantage (in part via counterforce) in the nuclear competition both for general deterrence and alliance cohesion benefits, as policymakers perceived a delicate nuclear balance where counterforce competition bore real peacetime benefits.[8] U.S. policy and nuclear employment guidance have long espoused flexible response options which include—but are not limited to—counterforce and damage limitation in case deterrence fails.[9] Starting in 2001, the U.S. began implementing national missile defenses as well as regional defenses in Europe and Asia to deter nuclear armed rogue states. U.S. missile defense policy states that

[5] Wohlstetter 1959.
[6] Wohlstetter 1959; Kahn 1965.
[7] Kahn 1965, p. 136-7.
[8] Green 2020, p. 172-3, 226-7, 247; Green and Long 2017b; Green and Long 2020.
[9] Lawrence Livermore National Laboratory 2023, p. 26-32.

national missile defenses are designed and intended to counter threats from North Korea and Iran, but the U.S. retains the right to defend its self against attacks from any source.[10] Robert Powell did seminal work in 2003 to formally model the deterrence logic of a stronger power like the U.S. fielding missile defenses with known efficacy against a lesser power like North Korea, highlighting the deterrence benefits and risks of the U.S. missile defense policy change.[11]

But since then, the major powers have gone even further to pursue counterforce policies for advantage, expanding the scope of the competition beyond Powell's concept of a one-sided competition of missile defense capabilities with known efficacy. Since 2017, U.S. Defense Department policy documents emphasize that missile defense is becoming more comprehensive by developing capabilities and authorities to intercept missiles in all phases of flight, and if deterrence fails, to also be able to attack missiles before they are launched with counterforce capabilities.[12] Even further, the U.S.'s major power rivals are also advancing their own counterforce capabilities.[13] Recent scholarship has highlighted that both the U.S. and China use non-traditional damage limitation tools for deterrence, including cyber, counter-space, and other capabilities which have deeply uncertain efficacy.[14] But scholars have not yet addressed this new dynamic of a dyadic, two-party competition in missile defenses, counterforce, and other damage limitation capabilities which inject significant uncertainty in the calculus over the balance of power.[15]

Some scholars have highlighted the applicability of deterrence theory to understanding

[10] U.S. Defense Department 2019, U.S. Defense Department 2022.
[11] Powell, 2003.
[12] U.S. Defense Department 2019; U.S. Defense Department 2017.
[13] Lieber and Press 2020, p. 108-110, 117-118. Lieber and Press 2017.
[14] Gartzke and Lindsay 2017a; Gartzke and Lindsay 2017b.
[15] Cunningham 2021.

how damage limitation could be used in a crisis, positing that states with superior damage limitation also are more able to run risks in crisis.[16] However, most of the recent formal scholarly contributions on nuclear deterrence focus on how crisis bargaining outcomes are driven by either superiority in numbers of weapons or in terms of nuclear weapons survivability, but not damage limitation capabilities which have properties that make their effectiveness—and their strategic impact—much harder to assess.[17] Other recent formal work has evaluated the utility of stronger powers' using counterforce capabilities short of war to improve their bargaining position by degrading a weaker challenger's military capability.[18] While this work evaluates the utility of counterforce for general deterrence and the avoidance of war, there remains a gap in the literature on the benefits of counterforce and damage limitation for bargaining in a nuclear crisis.

In this paper we work to fill this gap in the literature by directly evaluating the utility of damage limitation capabilities within a crisis bargaining context, seeking to understand how these capabilities with inherently uncertain efficacy can convey advantage within a nuclear crisis. We extend the logic of the delicate nuclear balance school by adapting Powell's game theory model to situations where one or both players have significant but imperfect damage limitation—which could include both missile defenses and a myriad of counterforce capabilities— the efficacy of which exists as private information. We assume that two nuclear armed players find themselves in a brinksmanship crisis with significant political stakes and we model the behavior of the players under conditions of differing uncertainty about each other's disarming damage limitation capabilities. We also assume that a mounting crisis will be marked by uncertainty about those stakes and thus also by a competitive testing of resolve. In a crisis

[16] Green and Long 2017a, p. 196-199; Talmadge 2022, p. 26-30.
[17] Kroenig 2013; Kroenig 2021.
[18] Schram, 2021

between the U.S. and China, a central question would be whether Beijing believes it can break Washington's resolve to protect its allies by raising its expected costs of doing so, due to the presence and posture of their strategic forces and damage limitation capabilities. We describe how these capabilities relate to deterrence theory, and then we set up a game theoretic framework which incorporates the uncertainty that is characteristic of the efficacy of counterforce and missile defenses. Because these game theoretic equations are not analytically tractable, we use high performance computing to numerically solve for the game equilibria at thousands of unique points in the game's parameter space.

We begin the article by establishing our game theoretic approach. Game theory has a rich history in studying the strategic nature of nuclear deterrence, and it is especially useful here given the lack of a strong empirical record. Next, we review the body of contemporary work on game theory modeling of brinksmanship and nuclear deterrence in the shadow of missile defenses, and we then adapt these models to incorporate the characteristics of competition in damage limitation. We then lay out our game framework, which builds on a second-price, all-pay auction that Robert Powell developed but we add in the ability of both sides to limit damage under conditions of uncertainty.[19] We numerically approximate equilibrium solutions to the game using high-performance computing, and we show how our model captures different behavior than Powell's. We then evaluate the findings from our model in the case of the U.S.-China dyad using campaign analysis models and published values for both U.S. and Chinese capabilities from recent scholarship, specifically the work of Wu Riqiang.[20]

Our analysis shows that China's bargaining leverage is growing as it builds its arsenal

[19] Powell 2003.
[20] Riqiang 2020; Tecott Metz and Halterman 2021.

towards parity with the U.S., under the key assumption that the U.S. cannot threaten to use more than half of its nuclear force against China because it must retain at least an equal sized force to deter Russia in the aftermath of a putative U.S.-China conflict. We then use our model to evaluate the utility of different changes to the U.S. force posture, finding that options for qualitative advantage can significantly improve the U.S.'s crisis bargaining leverage while simultaneously limiting arms race spiral dynamics. We close with implications for theory, we review the limitations of our model, and make suggestions for future research.

Nuclear Deterrence Theory and Damage Limitation

Traditional nuclear deterrence theory applies to the scenario where two opposing states, armed with nuclear weapons and the associated confidence that they have a secure second-strike force, are engaged in a game of brinkmanship. While both parties wish to prevail in the crisis and are thus incentivized to stand firm, there is substantial risk in standing firm because the risk of escalation involving assured second-strike capabilities makes the potential costs very high. But the symmetric presence of second-strike weapons does not eliminate the existence of political disputes in the international system. Accordingly, when the two parties compete during a political crisis, they ultimately are engaged in nuclear brinkmanship, which is defined as the willingness to risk nuclear escalation to achieve their objectives. In the evolving crisis, the side willing to take greater risks is often the one with higher stakes in the standoff. Thus, this game gives the advantage to the side with higher stakes, as they can run greater risks and thereby have a greater ability to coerce.

But the introduction of new capabilities that weaken the viability of second-strike nuclear forces reduce the certainty of mutual annihilation, and thereby invites greater risk taking. Missile defense capabilities can accomplish this asymmetric reduction of risk. Powell studied

this idea, and he assumed that one side's missile defense efficacy is clearly known by both sides. Powell found that the side with greater missile defense capabilities has a known advantage in brinksmanship.[21] But Powell's assumption on the transparency of missile defense efficacy is not realistic, as missile defenses are becoming more advanced and layered, and as rivals adjust by introducing countermeasures. Powell also did not include counterforce capabilities in his model, but they similarly impact the risks and costs of mutual destruction as their efficacy is also hard to judge. Because there may be substantial uncertainty around the effectiveness of both missile defense and counterforce capabilities, the reliability of the technologies involved, and both states' beliefs around said effectiveness, these ambiguities should impact brinkmanship outcomes. However, it is not a priori clear how such uncertainty around counterforce and its combined impact with missile defense technologies will change these outcomes. By explicitly incorporating uncertainty around these technologies into a brinkmanship model, we aim to shed new light on this phenomenon.

MAD, THE CREDIBILITY PROBLEM, AND MISSILE DEFENSE

Deterrence theory in the 1950's and 1960's focused on the implications of the U.S. and the Soviet Union obtaining second-strike nuclear arsenals, which meant that both sides could absorb a first strike without losing the ability to retaliate. But even once these second-strike forces were in place, political tensions between the two sides remained. As a result, they had to rely on deterrence to keep the other from attacking in a political crisis, but they also faced incentives to use coercive pressure against one another. The tension between the need to uphold deterrence and the simultaneous incentives to find coercive leverage result in what Powell calls

[21] Powell 2003, p. 89-91.

the credibility problem at the center of nuclear deterrence theory.[22]

One way that states can solve this credibility problem in the face of mutual second-strike capabilities is through brinksmanship. If there is a strong asymmetry of stakes in such a game, the state with more at stake can afford to stand firm and can do so by running risks of inadvertent or accidental escalation. Schelling described this as the condition in which neither side can credibly threaten deliberate nuclear attack, but where one side may credibly make "threats which leave something to chance."[23] In this contest, the state with higher stakes can out-bargain the other—via brinksmanship—by running greater risks of inadvertent or accidental escalation which thereby signal greater stakes and resolve.

Academics developed and applied these models to the U.S.-Soviet relationship during the Cold War. But after the fall of the Soviet Union, the prospect of mutual vulnerability with nuclear "rogues" like North Korea and Iran left U.S. policymakers uneasy, as the prospect of these states acquiring nuclear weapons would leave Washington in a position where it could be out-bargained via brinksmanship in a regional crisis where the U.S. would be extending its nuclear umbrella to regional allies.[24] In this context, U.S. policymakers have extended its nuclear umbrella to allies as one key means of reducing the incentive for allied states to develop their own nuclear weapons.[25] Following this logic, the U.S. withdrew from the anti-ballistic missile treaty in 2001 and deployed limited national missile defenses to counter rogue nuclear states with limited nuclear arsenals. Powell applied game theory to this situation to understand the value of U.S. national missile defenses in crisis, using a second-price, all-pay auction model to describe

[22] Powell 2003, p. 89-91.
[23] Schelling 1960, p. 187.
[24] Rumsfeld Commission 2001, p.2
[25] Gavin 2017.

how the effectiveness of the U.S. missile defense system shapes brinksmanship and risk taking—as well as how it impacts the risks of first-strike stability.[26] Powell's model predicts that missile defense capabilities with clear efficacy will increase the U.S.'s bargaining position in brinksmanship despite a disadvantage in stakes—but while marginally increasing risk of disaster as missile defense effectiveness grows high because the U.S. is more likely to intervene with more effective defenses.[27] We improve and extend Powell's model by introducing uncertainty around damage limitation efficacy—thereby capturing both advanced missile defenses and counterforce in our model—which complicates crisis entry conditions and subsequent crisis bargaining. We then also extend the model to understand how bargaining changes when both rivals in a dyad have imperfect and uncertain damage limitation capabilities, like those we see in the current and future security environment.

MODELING DAMAGE LIMITATION WITH GAME THEORY

In this section we review the details of second-price, all-pay auction models that scholars have developed to describe nuclear brinksmanship, and we adapt these models to also incorporate counterforce capabilities that carry much more uncertainty around their perceived effectiveness. The second-price, all-pay auction model we employ is a type of "war of attrition" model in continuous time like those used by Henricks, Weiss and Wilson.[28] Barry Nalebuff first proposed an incomplete information version of the model as an approach to nuclear crisis bargaining, and Powell later modified and extended it.[29] These games are useful because they speak to Schelling's original solution to the commitment problem in nuclear brinksmanship.

---

[26] Powell 2003.
[27] Powell 2003, p. 110.
[28] Hendricks, Weiss, Wilson 1988.
[29] Nalebuff 1986; Powell 2003.

Namely, rational actors have a hard time credibly threatening the use of nuclear weapons when second-strike capabilities are robust, since the risks associated with a general nuclear exchange are simply too high. But players might be willing to demonstrate their resolve by progressively escalating the crisis and increasing the risk that things spiral out of control.

Both Nalebuff and Powell's approaches rest on the idea that a nuclear crisis can be modeled as a second-price, all-pay auction where bids are associated with the amount of risk a player is willing to incur in order to prevail. In this context risk is equated with the probability that the crisis spirals out of control resulting in a general nuclear exchange, which we will refer to as the probability of a general nuclear exchange. Whichever player bids the larger risk (or tolerates the highest probability of a general nuclear exchange) wins the auction, but both players must pay the price set by the losing player by experiencing the mutual probability of a general nuclear exchange. The reason for this is the inherent assumption that the crisis unfolds until the minimum risk level is achieved, at which point the player with the lower bid submits. Thus, even though the player with the higher bid prevails in the crisis, both players experience the risk set by the losing player's bid.

More formally, there are exactly three possible outcomes in the auction game: a) player 1 concedes while player 2 prevails, b) player 2 concedes while player 1 prevails, and c) the crisis spirals out of control resulting in a general nuclear exchange. We assume each of these outcomes is associated with a specific payoff for each player. The payoff to prevailing for player $i = 1,2$ is denoted by $w_i > 0$. The cost of conceding for player i is $s_i \geq 0$. The cost of a general nuclear exchange is $d_i > 0$. In general, we assume the costs and payoffs are unique to each player.

In Nalebuff's version of the game, each player's bid is the amount of time they are willing to stay in the nuclear brinkmanship game—but time is only an instrument for risk—as he

assumes that staying in the game longer involves incurring progressively higher amounts of risk.[30] In Powell's missile defense model the players explicitly choose the risks they will incur rather than the amount of time they are willing to remain in the crisis. Following Powell's formulation each player's strategy is to choose the maximum risk (probability of a general nuclear exchange) they are willing to incur in order to prevail in the crisis.

Nuclear brinkmanship is a contest of resolve. Whichever side has more resolve is likely to tolerate higher risks, thus outlasting the less resolute player who is forced to acquiesce. Powell defines resolve as the greatest amount of risk a player is willing to incur in order to succeed if she knew that it would guarantee her success. This quantity is fundamentally connected to the stakes of the crisis, namely the costs and payoffs of the auction. With this definition, Powell explicitly defines the resolve of a player in terms of the costs and payoffs of the auction model. Letting R_i denote the resolve of player i, and given the above definition, it can be shown that $R_i = \frac{w_i+s_i}{w_i+d_i}$. Without loss of generality, we rescale the parameters in such a way that $s_i = 0$, giving us the definition of resolve as $R_i = \frac{w_i}{w_i+d_i}$.

With this formulation we see how the resolve of a player is a function of both the player's stakes in the crisis, w_i, and the player's potential cost of ruin, d_i. Furthermore, we see how any uncertainty about an opponent's stakes or potential costs of nuclear war directly leads to uncertainty about their resolve. Such uncertainty is fundamental. If the resolve of each player were known to her opponent, then there would be no crisis: The player with the lower resolve would submit immediately to minimize the risk of a general nuclear exchange. Thus, uncertainty around resolve is a necessity in true brinkmanship crises. To finalize the formulation of this

[30] Nalebuff 1986.

game we must account for such uncertainties. Specifically, we assume the stakes and costs, w_i and d_i respectively, are private information for player i.

As Nalebuff shows, bidding strategies are uniquely determined by the resolve $R_i = \frac{w_i}{w_i+d_i}$, so it suffices to consider an incomplete information game where the private information is the player's resolve. A strategy in such a game is a function $r_i(R_i)$ which maps a player's resolve (her private information R_i) to the risk (the probability of a general nuclear exchange r_i) she is willing to tolerate when her realized resolve is R_i. With resolve characterizing a player's type, common knowledge probability distributions representing beliefs about opponent resolve fully specify the incomplete information brinkmanship game. Given player beliefs about opponent resolve we can write down the expected payoff for player i with resolve R_i when choosing a risk level r. Given these expected payoffs, it is possible to determine equilibrium conditions, which when solved lead to a pair of equilibrium strategies $(r_1^*(R_1), r_2^*(R_2))$ which map player types (resolve, R_i) to equilibrium strategies (risks, r_i^*). Details of these equilibrium conditions can be found in the technical Appendix.

Powell built on the above model in two ways. The first, and most relevant extension for our purposes, is that he assumed that one player (the U.S., in Powell's game) possesses a national missile defense system capable of stopping a nuclear attack with probability of success e. If player i has the national missile defense capability, then the cost of a nuclear attack is reduced to the expected cost $(1 - e)d_i$. Powell further extended this model to include preliminary actions made by both the U.S. and a nuclear armed rogue state. The rogue state makes some initial move to initiate a conflict. The U.S. must decide whether to engage the rogue state, while the rogue state must decide whether to remain in the conflict if the U.S. engages. If the U.S. engages

and the rogue state remains, then a nuclear brinkmanship auction game similar to the one presented above unfolds.

We build on the Nalebuff-Powell models in several ways. Powell's model assumes that both national missile defense effectiveness e and the cost of a nuclear attack d_i are common knowledge, so there is no uncertainty around these values and the general balance of power. Instead, there is only uncertainty around the balance of resolve, as expressed by the winning payoff w_i. The uncertainty on the winning payoff then defines the uncertainty on the players' resolve. We believe that the lack of uncertainty around the efficacy of missile defenses and counterforce is a key omission of Powell's model, and this shortcoming needs to be amended to be able to extend the model to the contemporary security environment.

Moreover, we are interested in including nuclear dyads where both parties possess some form of damage limitation capability, as China is also working to develop its own damage limitation capabilities. Letting e_i denote the effectiveness for player i's damage limitation capability, we seek to extend this model to the case where both players possess a capability to deny each other's ability to use their strategic forces. We assume that the effectiveness of these capabilities is private information to the possessor of the capability.

With these modifications, all game payoffs (d_i, s_i, w_i) as well as the comprehensive damage limitation effectiveness (e_i) are private information. We then introduce a flexible distribution over player types to allow for more robust modeling of player beliefs. As a result of these extensions, the equilibrium solutions are no longer available in closed form. As such, we must rely on numerical solutions to the first-order equilibrium conditions. In order to adequately explore the entire parameter space of the games we leverage high-performance computing to solve the equilibrium conditions at thousands of points in the game's parameter space. While we

could simply explore the parameter space using comparative statics for a handful of cases, here we are able to show a much more complete picture of how bargaining works as a nearly continuous game space of nuclear force balance conditions between rivals.

One limit to our model is that it assumes that the players are already in a situation of nuclear crisis brinksmanship. This is the case because the equilibrium crisis entry conditions of Powell's model do not admit unique solutions under our more general modeling framework without further assumptions on damage limitation in conventional conflicts. As we only seek here to understand whether and how damage limitation can be used to manipulate risk in a nuclear crisis, we assume a nuclear brinkmanship crisis has already been instigated, ignoring the preliminary entry decisions considered by Powell. Instead, we wish to explore the equilibrium risk bidding strategies of the players when both sides have a damage limitation capability of uncertain effectiveness. Even though we do not consider the preliminary decisions about entering into the crisis, we note that the uncertainty in the balance of power that is inherent in our model provides a second possible avenue that could facilitate entry into a crisis, adding it to the uncertainty inherent in the players' resolve.

UNCERTAINTY IN DAMAGE LIMITATION EFFECTIVENESS

While uncertainty is a critical feature of counterforce and advanced missile defense capabilities, little prior research has assessed how this uncertainty can drive consequences in brinksmanship scenarios. Prior work by Braden Soper evaluated how incorporating the uncertainties of damage limitation capabilities into the game of Chicken changes the nature of strategic stability, namely by creating large regions in the game's parameter space where standing firm becomes a dominant strategy for players with sufficiently effective damage

limitation capabilities.[31] One shortcoming of this modeling approach is that the resulting games are all static matrix games and do not explicitly address the issue of a player's resolve in what is a naturally dynamic process. We thus seek to combine the uncertain damage limitation capabilities from Soper's models with the second-price, all-pay auction models of Nalebuff and Powell. We do this in two ways. First, we consider the value $R_i^* = \frac{w_i}{w_i+(1-e_i)d_i}$, which we will refer to as the *effective resolve* of player i, as the player type. Second, we introduce a joint probability distribution over damage limitation effectiveness e_i and the value $R_i = \frac{w_i}{w_i+d_i}$, which we refer to as the *baseline resolve* of player i. This joint probability distribution characterizes their opponent's beliefs over all their private information, which is completely characterized by the player's effective resolve R_i^*. Note that this follows directly from the analysis of Nalebuff by simply replacing d_i with $(1-e_i)d_i$.

To fully specify this model, we need to specify probability distributions over the effective resolve R_i^* that characterizes an opponent's beliefs. Because we want to understand the effects of uncertainty in damage limitation, we directly model the marginal distribution of the damage limitation effectiveness e_i. We can then fully specify the distribution on effective resolve R_i^* by noting the following relation between effective resolve, baseline resolve and effectiveness.

$$R_i^* = \frac{R_i}{R_i + (1-e_i)(1-R_i)}$$

Using this relation, we can specify distributions over baseline resolve R_i and damage limitation effectiveness e_i and then derive the distribution over effective resolve R_i^*. Because both baseline resolve and effectiveness can be interpreted as probabilities, we choose beta

[31] Soper 2019, p. 470-479.

distributions to model all uncertainty around these values. The beta distribution is a flexible distribution for modeling random probabilities, ratios, or proportions. It can thus model a wide range of beliefs held about damage limitation effectiveness as well as baseline resolve. The benefit of this model is that it is quite flexible as a distribution over the beliefs of an opponent's effective resolve. Moreover, by specifying distributions over damage limitation effectiveness and baseline resolve, we can isolate the effects of uncertainty on these two quantities. Intuitively, we have identified the *ex ante* levels of effective resolve for player i as a function of the stakes and damage limitation effectiveness. This is equivalent to modeling a player's expectation of how long it will remain in a crisis as a function of its cost tolerance and perceived weapons capabilities.

The drawback of this model is that closed-form analytical equilibrium solutions are not possible. This is in contrast to the modeling approach Powell used, which permits analytical solutions but prevents the same types of analyses to be performed here. Specifically, the distributions in Powell's model are characterized by only a single parameter each, thus it is impossible to isolate the first two moments of baseline resolve, let alone model effectiveness as private information. By contrast the distributions considered in this paper are each characterized by four parameters, allowing us to isolate the first two moments of both baseline resolve and effectiveness independently. We implement numerical solvers to find equilibrium solutions following the general first-order conditions for equilibria set out by Nalebuff. The next section presents several results from our computational analysis.

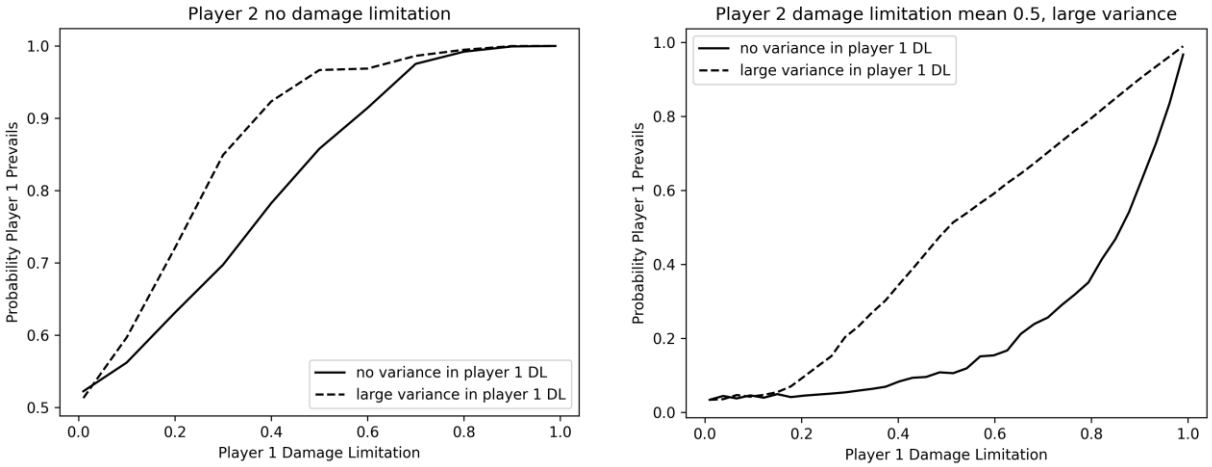
RESULTS

In this section we present the results of numerical solutions to the damage limitation brinkmanship game. The distributions over one player's beliefs about their opponent's effective resolve are defined by four parameters, two parameters for the baseline resolve and two parameters for the damage limitation effectiveness. A specific game is then defined by a total of eight parameters. For a given set of these eight parameters we can numerically approximate the Nash equilibrium (see the technical Appendix for more details on optimality conditions). Using high-performance computing resources, we numerically approximate the Nash equilibria at tens of thousands of points in the eight-dimensional parameter space in parallel. We present full results in the Appendix, while below we show representative results.

First, we compare our results to those of Powell's model, to determine whether in fact our model's differences produce different expected outcomes. To do this we fix player 2 as having no damage limitation capability while player 1 has varying damage limitation capabilities (as Powell did). We show results as player 1's damage limitation capability increases, under two scenarios: no variance (i.e., Powell's assumption that damage limitation efficacy is completely known to both players) and large variance (our assumption that damage limitation can have a lot of uncertainty around it). Conversely, we also consider the case where player 2 has a non-zero damage limitation capability with high uncertainty (mean 0.5, variance 0.083). Again, we consider what happens as player 1 increases their mean damage limitation under two scenarios: no uncertainty and high uncertainty. Figure 1 shows results comparing the probability that player 1 prevails under these different scenarios, highlighting that asymmetries in uncertainty around opponent damage limitation can be highly advantageous for the player that is more certain about their adversary's capabilities. Even when player 1 is completely certain that their adversary has

no damage limitation capabilities (left figure), player 1 can benefit from player 2 being uncertain about player 1's capabilities. The disadvantage player 1 faces when having their capabilities known to player 2 is even more pronounced when they themselves are highly uncertain about player 2's capabilities (shown in right figure as difference between the dotted and solid lines). These results are not surprising given the importance of information asymmetries in games of incomplete or imperfect information. But importantly, similar auction models such as Powell and Nalebuff's have neglected to address the effects of information asymmetries on game outcomes. We seek to remedy this with our modeling approach.

Figure 1. Results comparing the probability that player 1 prevails in a crisis as its damage limitation capability is increased under two scenarios. The solid line shows the scenario where there is no uncertainty in player 1's damage limitation efficacy, as Powell assumed. The dashed line shows the scenario where there is high uncertainty around player 1's damage limitation as measured by the variance. The figure on the left shows the scenario where the adversary (player 2) has no damage limitation. The figure on the right shows the scenario where the adversary has a mean damage limitation effectiveness of 0.5 with a high variance of 0.08.



Now that we have established that our model makes different predictions than Powell's we move to explore the model results more fully. A solution to the game is a pair of bidding strategies which are functions mapping a player's type (their effective resolve) to a risk tolerance. The risk tolerance is the probability a player is willing to accept that events "spiral out of control" and end in a general nuclear exchange. Given a profile of equilibrium bidding

strategies, we can compute various quantities of interest. The quantities shown in figures 2 through 7 are the *ex ante* expected probability of a general nuclear exchange and the *ex ante* probability of the U.S. prevailing. In the context of the game, the probability of a general nuclear exchange is the minimum bid between the two players in equilibrium, while the probability of the U.S. prevailing is the probability that the U.S. has the maximum bid in equilibrium. The *ex ante* expected utility for both players is provided in the Appendix, and we note that it closely tracks the probability of prevailing. In all figures, darker shades indicate higher values and lighter shades indicate lower values. Each figure contains solutions to games with fixed parameters for baseline resolve but varying parameters for damage limitation effectiveness. This allows us to isolate the effects of damage limitation effectiveness on brinkmanship behavior. Finally, examples of the equilibrium risk bidding strategies are shown in figures 8 and 9.

To better understand the effects of damage limitation effectiveness on nuclear brinkmanship crises, we systematically explore deviations in the mean and variance of both players' damage limitation effectiveness while keeping the mean and variance of the players' baseline resolve fixed. The mean of the damage limitation effectiveness distribution reflects the expected effectiveness of the opponent damage limitation capability, while the variance reflects the degree of uncertainty of the opponent damage limitation capability. Figure 2 through 4 show equilibrium results when we hold the mean damage limitation fixed at 0.5 for both players and unilaterally alter the variance of damage limitation for each player. The variance of the damage limitation increases along the x-axis for player 1 and along the y-axis for player 2. Figures 5 through 7 show equilibrium results when we hold the variance of damage limitation fixed at 0.01 for both players and unilaterally alter the mean of damage limitation for each player. Mean damage limitation increases along the x-axis for player 1 and along the y-axis for player 2. In all

figures the *ex ante* expected probability of general nuclear exchange is shown on the left and the *ex ante* probability of player1 prevailing is shown on the right.

The distributional parameters for baseline resolve are fixed at three different levels. Figures 2 and 5 show equilibrium results corresponding to a *symmetric, low-stakes crisis*, which we define as a crisis in which the payouts from winning are low compared to the costs of events spiraling out of control. In low-stakes crises, we assume both players have a distribution over their baseline resolve, R_i, that is a Beta(1,99), which has mean 0.01 and variance $9.8 \times 10^{-}5$. Figures 2 and 5 show equilibrium results corresponding to an *asymmetric stakes crisis*. In this case player 1 has a lower baseline resolve distributed according to Beta(1,99), while player 2 has a higher baseline resolve distributed according to Beta(1,9). Figures 4 and 7 show equilibrium results corresponding to a *symmetric high stakes crisis*, which we define as a crisis in which both players have a distribution over their baseline resolve, R_i, that is a Beta(1,9), which has mean 0.1 and variance 0.008. We next discuss why we chose these values, and we discuss these three sets of results (symmetric low stakes, asymmetric stakes, and symmetric high stakes) in more detail.

SYMMETRIC LOW STAKES GAMES

First, we consider the case that baseline resolve is relatively low (mean 0.01 and variance $9.8 \times 10^{-}5$). This corresponds to the idea that states fear nuclear holocaust far more than they value the potential payoffs of a political crisis, in which case the cost of events spiraling out of control are significantly higher, or alternatively that the political stakes are significantly lower. This set of games is meant to model the assumption that stakes must be extremely low in nuclear crises as the risk of annihilation greatly overwhelms the political costs or benefits of the dispute.

In figure 2, when mean expected damage limitation is held fixed, there is a clear effect on altering the variance of damage limitation. When the variance is mutually low or high for both players (which corresponds to low or high uncertainty about opponent damage limitation effectiveness, respectively), the probability of ruin roughly doubles (but is still fairly low at 0.5%) while the probability of either player prevailing is roughly the same (lower left and upper right corner). Conversely, when the degree of variance is highly asymmetric (which corresponds to one player being much more uncertain about the effectiveness of their opponent's damage limitation capabilities) the probability of ruin is halved while the player with higher variance in damage limitation is much more likely to prevail (upper left and lower right corners). These results suggest that *ceteris paribus* higher degrees of uncertainty in damage limitation (as measured by statistical variance) provide a strategic advantage even in a low stakes crisis. This suggests that the uncertainty around damage limitation capability can indeed be a central factor in the outcome of low stakes crises. Moreover, the advantage in damage limitation uncertainty seen here translates into a reduced risk of ruin, or more likely that the game that does not lead to nuclear war. However, if both players achieve a similar and very high degree of capability in damage limitation uncertainty this will result in a higher risk of ruin as the game becomes akin to a prisoners' dilemma as the conditions of nuclear deterrence weaken considerably.

In the other low-stakes game results shown in figure 5, when the variance of damage limitation is held fixed and we alter the mean, the mean damage limitation must be mutually very high for both players (i.e., both players believe that their opponent's damage limitation is over 90% effective) for there to be an appreciable increase in the probability of ruin. However, the probability of player 1 prevailing gradually increases (decreases) with even a moderate increase (decrease) in her mean damage limitation effectiveness. This suggests that even slightly higher

expected damage limitation provides a strategic advantage in a low-stakes crisis even under extreme assumptions about the fear and costs of nuclear annihilation. So, in the framework of our game, small damage limitation advantages provide meaningful crisis bargaining benefits even when the costs of ruin are extremely high, and these strategic advantages do not necessarily translate into a higher risk of ruin. But again, if both players achieve a high mean damage limitation capability this will result in a higher risk of ruin.

ASYMMETRIC STAKES GAMES

Next, we consider the impact of damage limitation on a set of cases where baseline resolve is asymmetric. This is meant to model scenarios in which two states enter a crisis where one state has higher political stakes in the outcome compared to the other. This should reflect the set of cases where the U.S. is projecting power far from its shores to support a weaker ally or partner state, and in particular when the U.S. is not a treaty ally with the weaker partner. A recent empirical study found that historically, nuclear advantages—including via damage limitation— in asymmetric stakes crises were not useful for the more powerful actor, as the authors hypothesized that the less powerful state will never enter a crisis in the first place if the combined balance of power and resolve is not in their favor.[32] We cannot test this logic formally here because the hypothesis is conditional on crisis entry decisions, which we have not evaluated. Rather we assume here that states with asymmetric stakes have already entered a crisis, and we seek to find how the balance of power affects the hypothetical outcomes.

While the U.S.' direct political stake in these crises is low, Jervis has argued that in many cases the U.S. has comparable stakes to its adversaries if only because—as the main status quo

[32] Fanlo and Sukin 2023.

power in the international system—it has more stake in maintaining the status quo than its challengers have in prospective changes to the international system.[33] Similarly, Brad Roberts argues that in a serious East Asian crisis with China, the U.S.' stakes would be perceived to be much higher than many analysts realize because U.S. credibility as a security guarantor would put its entire alliance structure at stake.[34] As a result, in this case we consider the case that the two states have only modestly different baseline resolves with mean $E[R_1] = 0.01$ for the player 1 (the lower-stakes player) and mean $E[R_2] = 0.1$ for player 2 (the higher-stakes player).

We find that when mean damage limitation is fixed and variance is altered, changes in player 2's damage limitation variance have a minimal effect on the game's outcome when compared to the symmetric-stakes games (see figure 3). On the other hand, unilaterally altering the variance of player 1 (the player with a lower baseline resolve) changes the outcome much more dramatically. In particular, the maximum probability of ruin has more than doubled from the low-stakes games, which occurs when both players have high degrees of uncertainty around damage limitation. Note also that unilaterally increasing the low-stakes player's variance increases the probability of ruin no matter the degree of variance in damage limitation of the high-stakes player. Thus, even though unilaterally increasing the variance of the low-stakes player's damage limitation increases her probability of prevailing, this comes at the expense of increasing the probability of ruin. Similar results can be seen in figure 6 when the variance is held fixed and the mean is varied: Unilaterally increasing the mean damage limitation of the low-stakes player increases her probability of prevailing, but this comes at the expense of increasing the probability of ruin. Regardless, the results here suggest that damage limitation, and

[33] Jervis 1989.
[34] Roberts 2020.

uncertainty around its efficacy, can make a very significant difference even in asymmetric stakes crises. This demonstrates that our model is capturing unique behavior around capabilities above and beyond simple differences in resolve or stakes between the two players.

SYMMETRIC HIGH STAKES GAMES

Figures 4 and 7 show equilibrium results for symmetric high-stakes games (mean $E[R_i] = 0.1$ and variance $Var[R_i] = 0.008$). This corresponds to the scenario in which both states value the potential payoffs from the crisis to a much greater degree, thereby increasing their risk tolerance of a nuclear holocaust. So, in this case either the cost of events spiraling out of control are significantly lower, or conversely that the political stakes are significantly higher.

In figure 4, where the expected damage limitation is fixed and the variance changes, we see that when the variance is high for both players (i.e., high uncertainty for both players' damage limitation effectiveness), the probability of ruin is slightly higher while the probability of either player prevailing is the same (see upper right corner). Conversely, when the degree of variance is highly asymmetric (i.e., one player being much more uncertain about the effectiveness of their opponent's damage limitation capabilities) the probability of ruin is relatively low. In these cases, the player with higher damage limitation variance is more than twice as likely to prevail. Note that the range of values for the probability of ruin are very high at roughly 5-6%, which is more than an order of magnitude higher than in the symmetric low-stakes games. Also, in contrast to the low stakes games, we see that mutually low variance does not lead to the highest levels of risk (lower left corner). Again, these results suggest that *ceteris paribus* a unilateral increase in damage limitation uncertainty (as measured by statistical variance) provides a strategic advantage in a high stakes crisis. This is a robustness check on our

results from lower stakes crises, in that again we see a prominent role for damage limitation and uncertainty in crisis outcomes.

Figure 7 shows that when variance is held fixed and the mean is altered, and there is low mean damage limitation effectiveness, unilateral increases in mean damage limitation effectiveness are unsurprisingly advantageous to the player with the higher mean effectiveness. This can be seen by the monotonic changes in shading on the bottom (x) and left-most (y) axes. The *ex ante* expected probability of a general nuclear exchange (left panel) monotonically decreases along these axes, while both the *ex ante* probability of player 1 prevailing (right panel) increase with higher mean damage limitation effectiveness. This suggests that asymmetric damage limitation capabilities provide a significant bargaining advantage to the player with higher expected damage limitation effectiveness, thereby incentivizing the player with lower expected damage limitation effectiveness to submit earlier in the crisis and reducing the risk of a general nuclear exchange. However, as both players mutually increase their mean damage limitation, the risk of ruin increases, just as in the low stakes scenario. Again, our findings here are consistent with the ideas of Kahn and Wohlstetter around the delicate balance theory, that even slight differences in the balance of power can make a large difference in the outcomes of high stakes crises.

EQUILIBRIUM BIDDING STRATEGIES

To better understand what drives the game outcomes presented above, we can look at the actual equilibrium bidding strategies $r_i^*(R_i)$, the equilibrium functions mapping player types to risk tolerance. Figures 8 and 9 show several equilibrium bidding strategies for both players under different scenarios. Figure 8 shows multiple equilibrium bidding strategy profiles corresponding

to unilateral increases in the mean of player 1's damage limitation effectiveness while keeping all other game parameters fixed. Figure 9 shows multiple equilibrium bidding strategy profiles corresponding to unilateral increases in the uncertainty (variance) of player 1's damage limitation effectiveness while keeping all other parameters fixed.

We fix the distribution on baseline resolve at a Beta(1,2), which has a mean of 1/3. This is arguably a large risk tolerance in a nuclear brinkmanship game. However, we have chosen this risk profile to aid in visualizing the mechanisms driving the equilibrium results. First, looking at figure 8 we can clearly see the effect of unilateral increases in mean damage limitation. In the left panel, we see that player 1 is able to increase her bids at lower realized values of effective resolve and decrease her bids at higher realized values of effective resolve. In particular, if player 1 has an effective resolve that is below (above) 0.5 then as her mean damage limitation increases, she will increase (decrease) her bids. In this way having an advantage in mean damage limitation effectiveness allows the player to moderate her risk tolerance away from extreme bidding behavior. Even more striking is the effect this has on player 2, seen in the right panel: For any realized effective resolve for player 2, she unilaterally decreases her bids as the mean damage limitation of player 1 increases. In other words, player 2 is cowed by her belief in player 1's superior damage limitation capabilities.

Next, in figure 9 we see distinct behavior when there is asymmetric uncertainty about the two player's damage limitation capability, as measured by variance. First, changes in bidding behavior are less pronounced when the variance unilaterally changes. Nevertheless, we see that increases in variance of player 1's capability consistently increases the bids of player 1 and decreases the bids of player 2. Thus, asymmetric advantages both in mean damage limitation capability and in uncertainty (variance) about the adversary's capability directly emboldens the

advantaged player and simultaneously cows the disadvantaged one. And further, we find that asymmetric uncertainty about the other side's capability can similarly affect bidding behavior.

Figure 2. Results of a low stakes (baseline resolve, R_i) standoff, showing expected probability of general nuclear exchange (left) and expected probability of player1 prevailing (right) under equilibrium strategies showing changes in outcomes with increases in damage limitation uncertainty (variance) along x- and y-axes. Symmetric distributions on baseline resolve are Beta(1,99), which has mean 0.01 and variance $9.8 \times 10^{-}5$. Mean damage limitation effectiveness is fixed at 0.5 for both players.

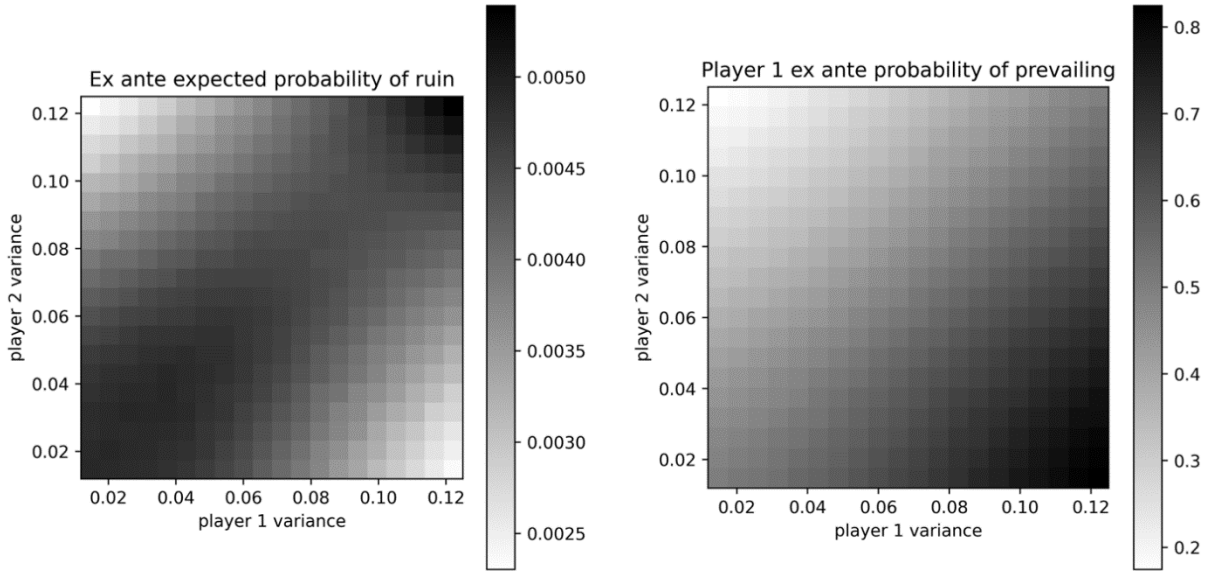


Figure 3. Results for an asymmetric stakes standoff with player 2 having higher baseline resolve, showing changes in outcomes with increases in damage limitation uncertainty (variance) along x- and y-axes. Asymmetric distributions on baseline resolve are Beta(1,99), which has mean 0.01 and variance $9.8 \times 10^{-}5$ for player 1, and for player 2 a Beta(1,9) which has mean 0.1 and variance 0.008. Mean damage limitation effectiveness is fixed at 0.5 for both players.

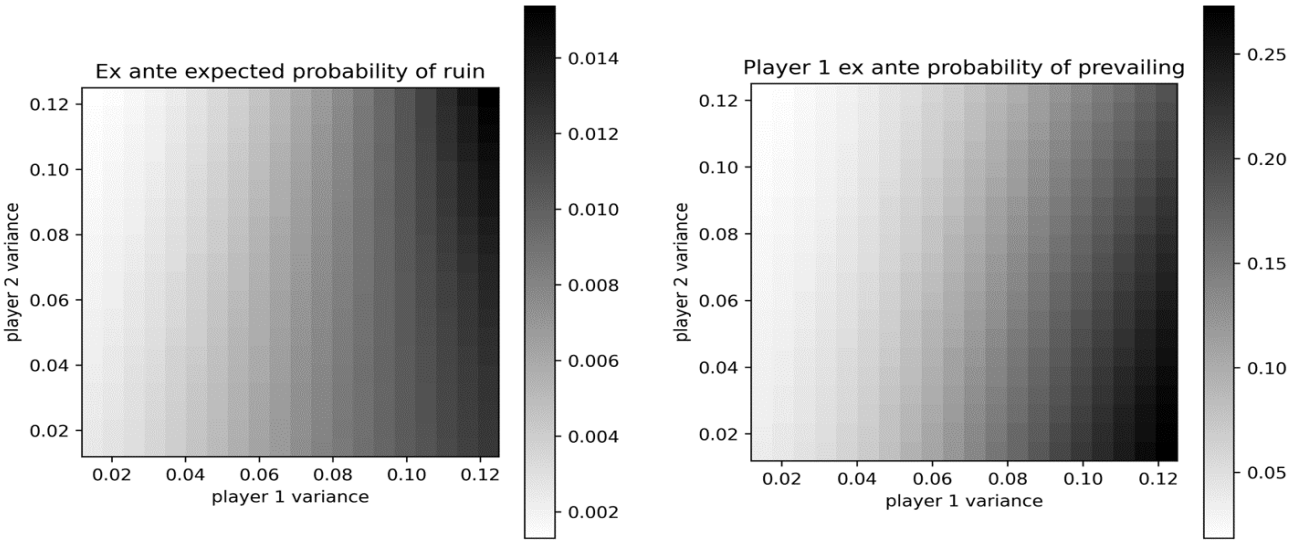


Figure 4. Results of a high stakes standoff showing changes in outcomes with increases in damage limitation variance changes along the x- and y-axes. Symmetric distributions on baseline resolve are Beta(1,9), which has mean 0.1 and variance 0.008. Mean damage limitation effectiveness is fixed at 0.5 for both players.

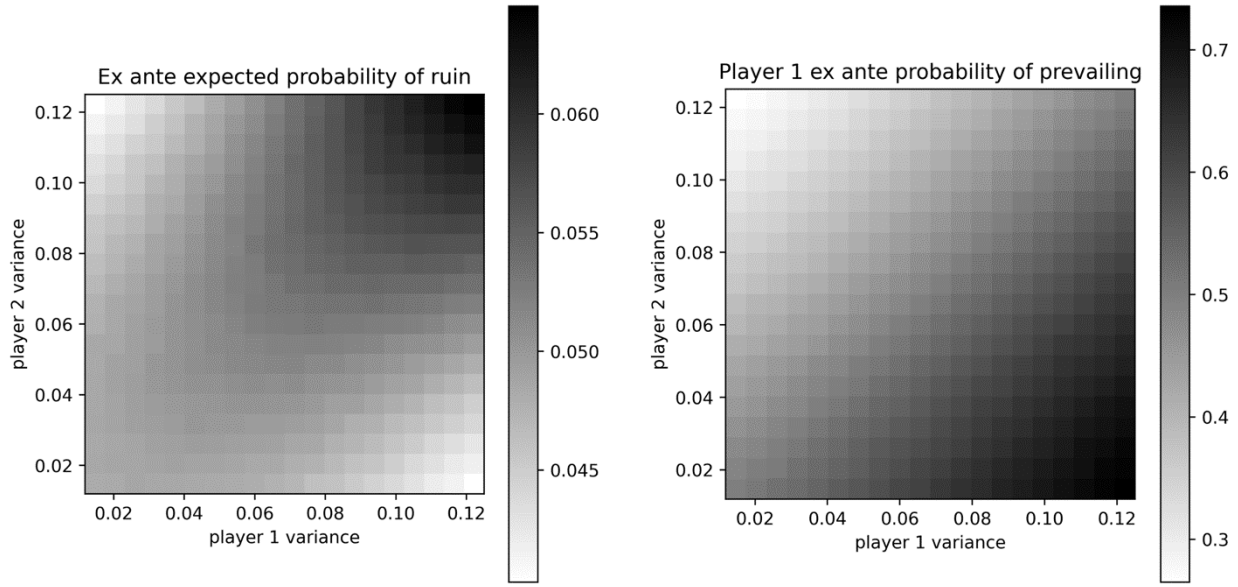


Figure 5. Results for a low stakes symmetric resolve crisis, showing changes in outcomes with increases in damage limitation efficacy (mean) along x- and y-axes. Symmetric distributions on baseline resolve are Beta(1,99), which has mean 0.01 and variance $9.8 \times 10^{-}5$. The damage limitation effectiveness variance is fixed at 0.01 for both.

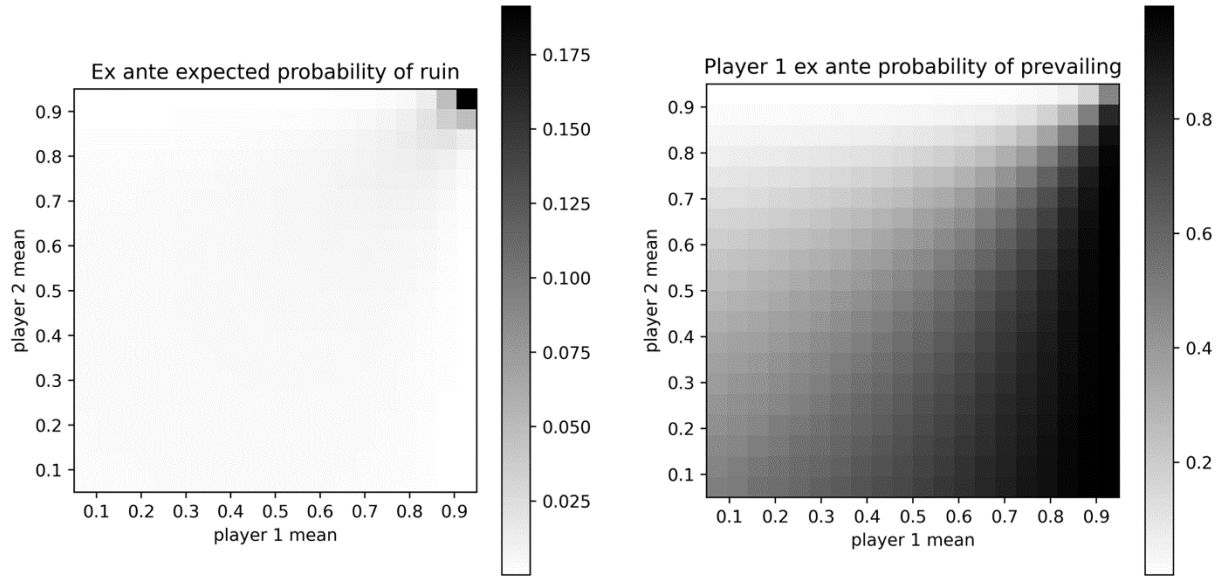


Figure 6. Results of an asymmetric stakes standoff, showing changes in outcomes with increases in damage limitation efficacy (mean) along x- and y-axes. Asymmetric distributions on baseline resolve are Beta(1,99), which has mean 0.01 and variance $9.8 \times 10^{-}5$ for player 1 and Beta(1,9), which has mean 0.1 and variance 0.008, for player 2. Thus player 2 has an advantage in baseline resolve. The variance of damage limitation effectiveness is fixed at 0.01 for both players.

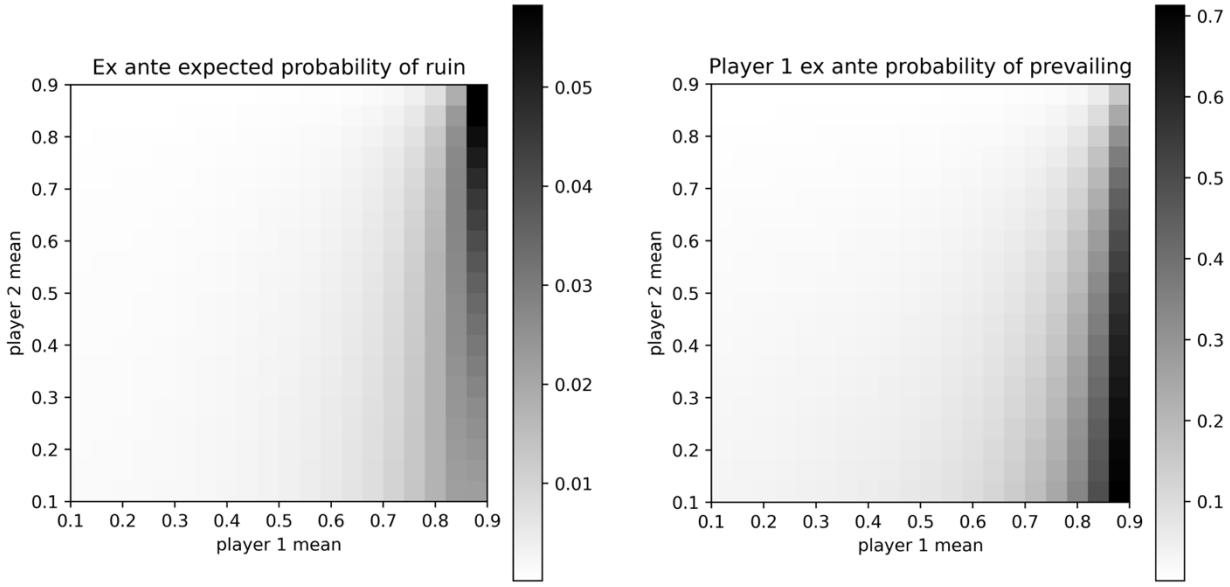


Figure 7. Results of a high stakes, symmetric baseline resolve standoff showing changes in outcomes with increases in damage limitation efficacy (mean) along x- and y-axes. Symmetric distributions on baseline resolve are Beta(1,9),

which has mean 0.1 and variance 0.008. The variance of damage limitation effectiveness is fixed at 0.01 for both players.

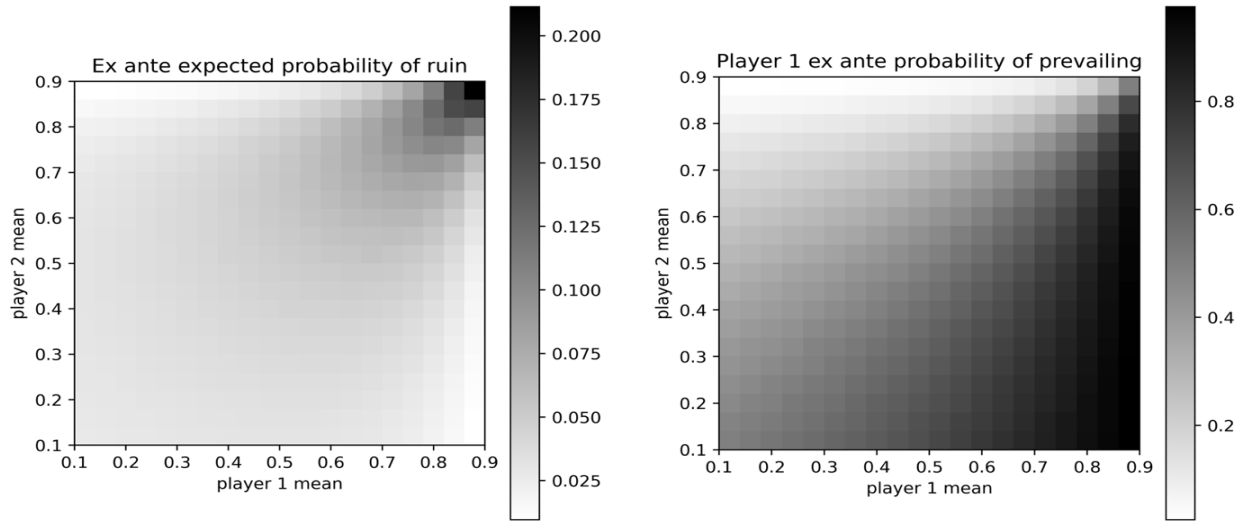


Figure 8. Changes in equilibrium risk bidding strategies with unilateral changes in player 1 mean damage limitation effectiveness. Stakes are high for both players, with baseline resolve distributed according to a Beta(1,2), which has mean 0.3333 and variance 0.0555, for both players. Player 1 mean damage limitation effectiveness is increased from 0.3 to 0.9 while the variance is fixed at 0.0153. The mean and variance of player 2's effectiveness is fixed at 0.3 and 0.0153, respectively.

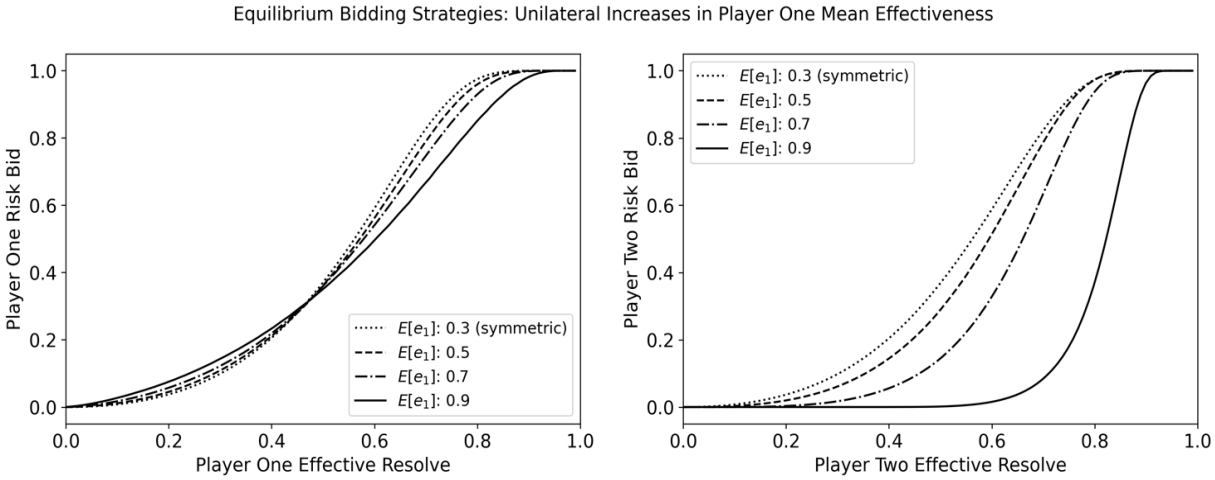
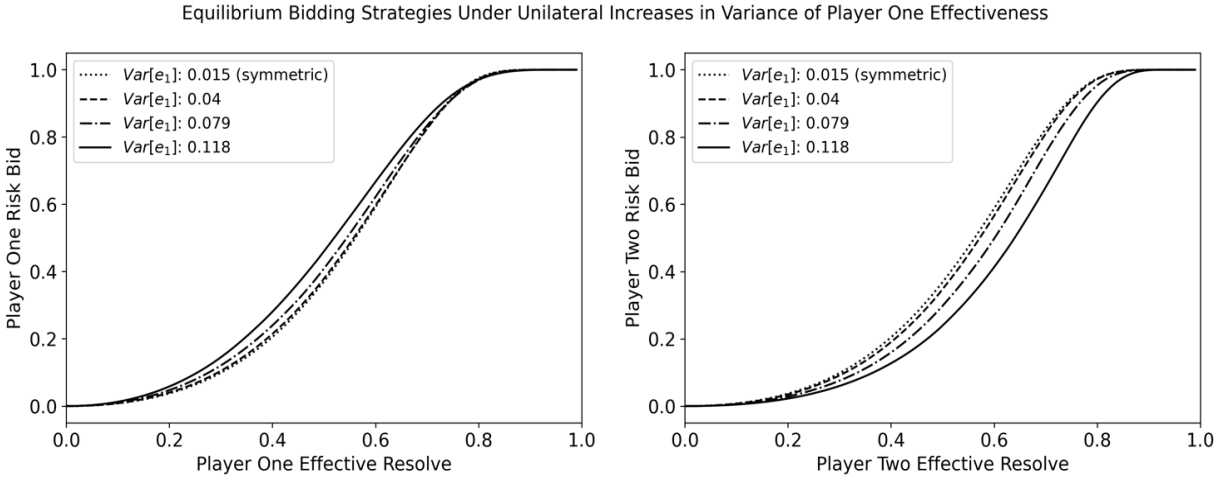


Figure 9. Changes in equilibrium risk bidding strategies with unilateral changes in uncertainty (variance) around player 1 damage limitation effectiveness. Stakes are high for both players, with baseline resolve distributed according to a Beta(1,2), which has mean 0.3333 and variance 0.0555, for both players. Player 1's variance on damage limitation effectiveness is increased from 0.0153 to 0.1181 while the mean is fixed at 0.3. The mean and variance of player 2's effectiveness is fixed at 0.3 and 0.0153, respectively.



Empirics

These results show that even a marginal advantage in damage limitation capabilities—either if it is known or suspected—should yield coercive leverage in nuclear brinkmanship crises even when resolve is asymmetric. To better illustrate the implications in today's international system, in this section we extend our model to the present and future security environment for the U.S. and China as Beijing grows its nuclear force. To do so, we examine the strategic force balance between the U.S. and China from 2010 to 2021, then we look ahead to the anticipated force balance in 2026 and 2036. We then use the expected results from a U.S.-China nuclear exchange that incorporates damage limitation— and which corresponds to the different values generated by our brinksmanship model. These results capture the expected dynamics of nuclear brinksmanship between the U.S. and China under competition in nuclear weapons and damage

limitation, based on an existing open source campaign model. They are just one assessment of how the balance of power would translate to a nuclear exchange, so they should not be considered authoritative.

We do not attempt here to go back into the history of nuclear crises to validate our model, but we have reviewed the existing literature to assess whether or not our model is consistent with research findings to date. Central to the literature on nuclear crises is the unresolved debate over what defines a "nuclear crisis," and that this definition in practice likely drives the results of empirical outcomes of whether or not nuclear weapons advantages impact crises.[35] Ultimately, while some studies have found that the stronger side—typically measured by numbers of nuclear weapons— is typically emboldened in a crisis, few have been able to establish if and how the weaker party is cowed (as our model predicts). Betts' seminal qualitative work on nuclear crises did find some systematic effect of the Soviets being cowed in Cold War nuclear crises due to their disadvantage in the balance of power, but we believe more specific study is merited especially as more evidence is becoming available about Soviet crisis decision making.[36]

While we do not have a rigorous validation of our model in past crises, we still want to know what the model suggests about the U.S. strategic bargaining capacity against its main conventional competitor—China—as it grows its arsenal from a minimum posture to one that is much closer to parity with the U.S. Our model's main utility is that it allows us to take the expected portion of each side's nuclear forces surviving in the face of their adversary's damage limitation capabilities, and we use these values to calculate expected crisis bargaining outcomes. In this section we will briefly summarize the predictions of this model for the U.S.-China dyad,

[35] Kroenig 2013, p.152-154; Fanlo and Sukin 2023; Green, Long, Bell, Macdonald 2019, p. 130–139; Kroenig 2018; Sechser and Furmann 2013; Sechser 2019.
[36] Betts 1987.

and we will update its parameters based on changes in the current and expected future U.S.-China force balance.

Estimating Damage Limitation Efficacy in the U.S.-China Dyad

Here we assess the expected changes in China's nuclear forces through 2036, and how these changes will impact prospective U.S. and Chinese strategic conflict using campaign analysis. We then map the results of the campaign analysis to our game theory model of brinksmanship and bargaining. The U.S. Defense Department expects China to field roughly 350 new intercontinental ballistic missiles (ICBMs) by 2026, and to expand to up to 1500 total weapons by 2036. Our assessment of how the three legs of China's nuclear force will evolve under these limits is taken from a recent study group report convened by Lawrence Livermore National Laboratory.[37] The Defense Department also recently assessed that China has begun implementing a launch under attack posture, enabled by an initial set of three missile warning satellites and a number of ground-based radars. They also note that the use of launch under attack could still be consistent with Beijing's long standing no first use policy.[38]

Scholars have recently used campaign analysis to evaluate the survivability of the pre-2020 Chinese arsenal against U.S. damage limitation capabilities. Wu Riqiang's analysis found that the U.S.'s counterforce and missile defense capabilities would effectively result in only a 90% chance that at most one weapon penetrates to reach the U.S. with the 2010 balance of forces, assuming the Chinese were in high alert status with a fully generated force.[39] Tecott Metz and Halterman subsequently replicated Riqiang's analysis, finding nearly identical results by

[37] Lawrence Livermore National Laboratory 2023.
[38] U.S. Defense Department 2023, p. 99-100.
[39] Riqiang 2020.

using a slightly different model.[40] Both studies showed that China can only count on perhaps a handful of weapons surviving to delivery with its 2020 force posture, with the introduction of mobile missiles in 2010 as being a seminal moment when the Chinese arsenal began to have a possibility of successfully striking the U.S. mainland.

We assume that the recently revealed 350 new Chinese silo-based ICBMs will be available by 2026. Figure 10 shows our results for the 2026 and 2036 exchanges, using the updated Chinese force balances for those years into the Riqiang Monte Carlo model. The U.S.'s overwhelming counterforce first strike and missile defense capability still limits China's ability to deliver more than 20 weapons in 2026, according to Riqiang's model. While the rest of China's arsenal will grow much more significantly between 2026 and 2036, Beijing's land based nuclear force will grow only slightly further. This should improve China's ability to deliver its ICBMs, perhaps by up to 25 weapons which is a significant improvement over 2010 or 2020, but still a small number of weapons surviving to target. Figure 11 shows results for the same force balances in 2026 and 2036, but adding a perfectly effective Chinese launch under attack (LUA) capability. Here we find that LUA enables China to be able to to deliver 150-300 weapons to the U.S., roughly increasing its ability to deliver damage by ten-fold. We note here that our campaign analysis model did not consider China's sea and air-based weapons, which are expected to constitute roughly half of China's arsenal in 2036. This is because the Riqiang model forces us to assume (as both he and Tecott Metz and Halterman did) that the U.S. can successfully target these weapon systems with conventional (non-nuclear) counter-air and anti-submarine warfare operations. Given data on reliability of U.S. conventional attacks, future analysis could update the Riqiang model to account for less than 100% efficacy in these

[40] Tecott Metz and Halterman 2021.

conventional operations. Absent such data, here we use the same assumptions as Riqiang, Tecott Metz and Halterman.

Figure 10. Histogram results of 100,000 Monte Carlo simulations of Chinese weapons surviving to target using Riqiang model with the U.S. Defense Department's anticipated future Chinese strategic forces in 2026 (left) and 2036 (right). Frequency on y-axis shows proportion of model results falling in each bin, and bins are defined by the number of Chinese weapons reaching the U.S.

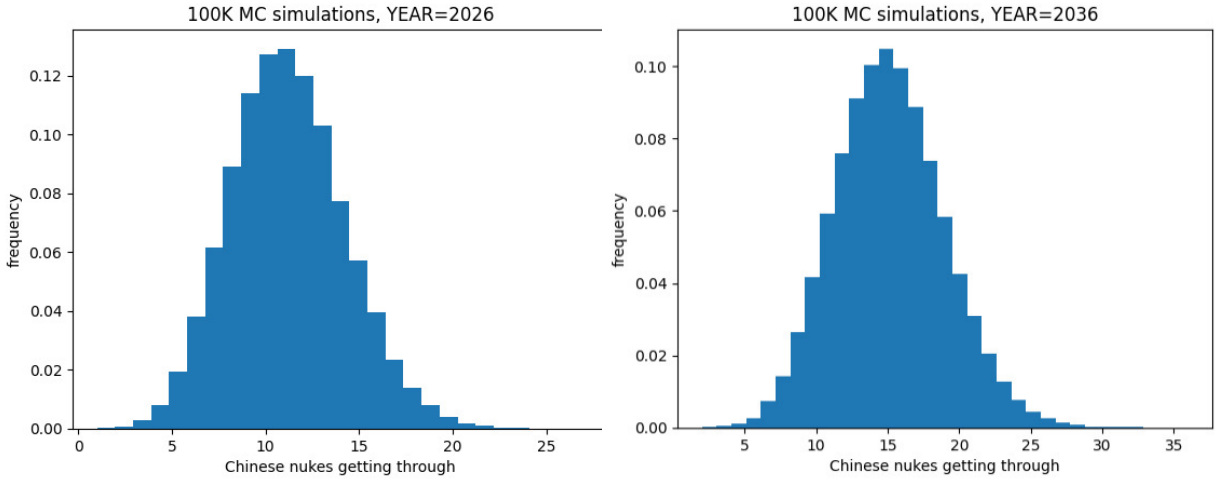
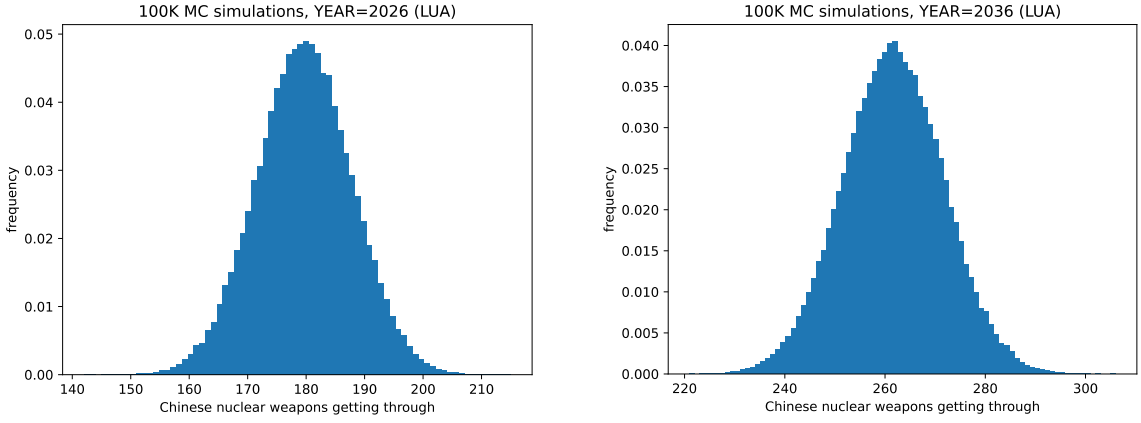


Figure 11. Histogram results of 100,000 Monte Carlo simulations of Chinese weapons surviving to target using Riqiang model with the U.S. Defense Department's anticipated future Chinese strategic forces with the addition of a launch under attack capability in 2026 (left) and 2036 (right). Frequency on y-axis shows proportion of model results falling in each bin, and bins are defined by the number of Chinese weapons reaching the U.S.



Estimating Brinksmanship Bargaining Outcomes in the U.S.-China Dyad

Next, we take the campaign analysis results and plug them into our game theoretic model

of bargaining. Details on how we translated campaign analysis outcomes to parameters in our model can be found in the appendix. We have two free parameters in our model. One parameter is the size of the disarming strike that the U.S. could deliver, which contains two key considerations. The first is how many U.S. weapons are needed to eliminate all Chinese silos, which in the Riqiang model is 1 per target, as the assumed single-shot-kill probability is 99.95%. The second consideration is how many weapons the U.S. would want to retain to maintain deterrence against Russia after the conflict. As Caitlin Talmadge has pointed out, in the Cold War U.S. planners took this into consideration with regard to planning nuclear operations against the Soviet Union, with the idea in mind that the U.S. would need to retain a certain minimum number of weapons after a war with the Soviets in order to maintain deterrence with China.[41] U.S. decision makers are likely to extend this same logic to current and future U.S. military planning as well. We assume there will be no major changes to New START levels of U.S. nuclear forces through 2026 and 2036, so we believe a credible U.S. first strike threat to China would be limited to likely under 750 weapons or roughly half of the total deployed force in order to retain a sizable deterrent against Russia.

The second free parameter in our model is the efficacy of China's damage limitation capabilities against the U.S. arsenal, which we assess will remain fairly low, only being capable of defeating a small portion of the significantly larger U.S. nuclear force. This is because of the small number of weapons reaching the U.S. mainland due to an assumed U.S. first strike. We make this assumption because China scholars anticipate Beijing to continue its no first use policy regardless of the U.S.' advantage in its nuclear and damage limitation postures, as it has held to

[41] Talmadge 2022.

this policy since 1964 and effectively keeps the burden of nuclear escalation on the U.S.[42] As a result, our model assumes that China will not be able to muster a significant damage limitation capability in the 2036 timeframe due to the U.S. advantage in anti-submarine warfare, and the lack of evidence of sufficient Chinese missile accuracy to conduct counterforce via a significant first strike against the U.S.' 400 silos. However, we leave this as a free parameter in our analysis—we evaluate Chinese damage limitation efficacy of 10%, 30%, 50%, and 70%— as the amount of damage limitation capability that China can field between now and 2036 is likely low but unknown.

We also must assume that there would be differences in the two sides' stakes. Given that we are assuming raids of large sizes, on the order of many hundreds of U.S. weapons, we selected distributions on stakes that resulted in small baseline resolves (i.e., payouts of the model expressed as *w* are still quite low relative to the high losses of *d* as we did not want high baseline resolve values, because this breaks the basic assumptions of the nuclear era). Finally, we selected gamma distributions around stakes (*w*) due to the inherent uncertainty between the players about each other's resolve in a crisis. Together the distributions around stakes and capability set our baseline resolve values. For the U.S. we assume $w \sim \text{Gamma}(1,4)$, total damage *d* varies by year with the total number of Chinese ICBMs (44 weapons 2010, 110 in 2021, 374 in 2026, 536 in 2036). For China we assume $w \sim \text{Gamma}(1,4)$ for symmetric stakes games and $w \sim \text{Gamma}(1,20)$ for asymmetric stakes, in other words that China's gains for winning are 5x more than the U.S. in the asymmetric case. Damage limitation results change every year based on the change in Chinese order of battle and our model assumptions.

Our results are shown in tables 1 through 4. We find a significant reduction in the U.S.

[42] Cunningham and Fravel 2015.

bargaining position with China through 2036, especially if China is able to field an effective launch under attack capability. Indeed, the two key factors that lead to Chinese leverage in this analysis is China convincing itself and the U.S. that it has asymmetric stakes, and its implementation of launch under attack. The models show that there is a low but growing possibility of a general nuclear exchange especially when both sides have high uncertainty about each other's damage limitation capabilities. We find that regardless of assumptions, China's growing nuclear force buys it coercive capability but at the cost of an increase in the probability of ruin. The U.S. becomes unlikely to win crises in 2036 when stakes are asymmetrical and where China's launch under attack capability is effective.

However, if the stakes are symmetrical, China's likelihood of winning a future crisis against the U.S. only exceeds a 50% chance in cases where it develops a significant (>30% effective) damage limitation capability of its own. If China credibly holds to its no first use policy, a significant damage limitation capability would require a major advance in its ability to conduct anti-submarine warfare against U.S. strategic missile submarines, which carry over 50% of the U.S. deployed nuclear force.[43] Though unlikely, China could also achieve a significant damage limitation capability via abandoning no first use and increasing the accuracy of its long range missiles for precision strike against the U.S.'s 400 silos. Our findings show that China's ability to gain strategic coercive advantage via damage limitation capability can occur in combination with either high variance damage limitation capability (i.e., via keeping the U.S. in the dark about its true capability) or via convincing the U.S. that Beijing has much greater gains to be won in the crisis outcome.

Table 1. Chances of the U.S. prevailing in a symmetric stakes crisis, with low variance Chinese damage limitation capability, assuming different mean values in columns. Each cell shows U.S. probability of winning the crisis in the top number, while the bottom number in each cell shows the concomitant probability of ruin. Results derived using

[43] U.S. Navy Website 2023.

Riqiang model with assessed or anticipated U.S.-China force balances. Last row shows results an effective Chinese launch under attack capability in 2036.

Chinese Damage Limitation Efficacy

Year	10%	30%	50%	70%
2010	99.83% 0.01%	99.75% 0.01%	99.72% 0.02%	99.56% 0.04%
2021	99.53% 0.02%	99.42% 0.03%	99.20% 0.05%	98.53% 0.11%
2026	99.41% 0.03%	99.27% 0.04%	98.97% 0.07%	98.09% 0.13%
2036	99.28% 0.04%	99.06% 0.05%	98.68% 0.09%	97.44% 0.17%
2036 w/ LUA	83.6% 0.2%	75.77%. 0.27%	63.87% 0.32%	44.37% 0.40%

Table 2. Chances of the U.S. prevailing in a symmetric stakes crisis, with high variance Chinese damage limitation capability (i.e. U.S. has less insight on Chinese capabilities). Same format and assumptions as Table 1.

Chinese Damage Limitation Efficacy

Year	10%	30%	50%	70%
2010	99.82% 0.01%	99.33% 0.05%	95.45% 0.46%	93.33% 0.98%
2021	99.47% 0.02%	98.47% 0.08%	91.47% 0.49%	86.55% 0.97%
2026	99.38% 0.03%	98.11% 0.10%	90.18% 0.50%	84.32% 0.95%
2036	99.21% 0.04%	97.63% 0.11%	88.66% 0.51%	81.46% 0.93%
2036 w/ LUA	83.32% 0.23%	69.75% 0.28%	38.66% 0.26%	22.08% 0.26%

Table 3. Chances of the U.S. prevailing in an asymmetric stakes crisis (China has higher stakes), with low variance Chinese damage limitation capability (i.e. U.S. has high insight on Chinese capabilities). Same format and assumptions as Table 1.

Chinese Damage Limitation Efficacy

Year	10%	30%	50%	70%
2010	99.28% 0.10%	99.12% 0.14%	98.77% 0.24%	98.07% 0.54%
2021	97.59% 0.24%	96.88% 0.36%	95.50% 0.61%	91.78% 1.38%

2026	96.73% 0.29%	95.74% 0.44%	93.74% 0.73%	88.39% 1.60%
2036	95.43% 0.35%	93.90% 0.53%	90.95% 0.89%	83.35% 1.84%
2036 w/ LUA	28.38% 0.43%	21.12% 0.42%	14.04% 0.38%	7.48% 0.28%

Table 4. Chances of the U.S. prevailing in an asymmetric stakes crisis (China has higher stakes), with high variance Chinese damage limitation capability (i.e. U.S. has less insight on Chinese capabilities). Same format and assumptions as Table 1.

Chinese Damage Limitation Efficacy

Year	10%	30%	50%	70%
2010	99.25% 0.10%	98.02% 0.36%	91.44% 1.93%	86.64% 3.81%
2021	97.54% 0.24%	94.26% 0.65%	80.79% 2.08%	69.55% 3.71%
2026	96.59% 0.30%	92.40% 0.73%	76.83% 2.07%	64.07% 3.48%
2036	95.27% 0.37%	89.96% 0.82%	71.87% 1.99%	56.67% 3.17%
2036 w/ LUA	28.02% 0.43%	17.95% 0.37%	7.31% 0.20%	3.57% 0.14%

U.S. Competitive Measures

Last, we evaluate how changes in the U.S.'s force posture impacts its utility function, in order to understand how different combinations of possible future capabilities lead to improvements in the U.S.'s bargaining position. The figures below show this with results from 2036, and with both symmetric (left figures) and asymmetric (right figures) resolve cases. First and foremost, we find that when the U.S. has convinced itself and China that its stakes are symmetric, the U.S.'s bargaining leverage is much greater. Thus, measures to tie U.S. credibility to the outcome of the crisis, particularly in its willingness to support other allies and partners, would help buy crisis bargaining leverage. Demonstrating the importance of a specific crisis to the credibility of the U.S.' broader network of alliances and partnerships would demonstrate that

it has more at stake than just the immediate stakes in question. This logic could also be a unique advantage to the U.S. side as China lacks similar such alliances and partnerships.

In Figure 12 we show our analysis of the trade in U.S. utility between changing the potential raid size the U.S. can threaten (or amount *d* the U.S. can threaten against China) versus small changes in efficacy (*e*) of its damage limitation. Here we see that very small changes in damage limitation efficacy, say 5-10%, which could be realized via improvements in missile defense, missile accuracy, or other cross domain means, is equivalent to changes of say 500 additional U.S. deployed weapons with which it can threaten larger disarming strikes. These large changes in threatened strikes could be achieved via a larger nuclear force or perhaps just with more significant nuclear weapons upload capability. While there are no planned changes to the U.S. deployed arsenal, it is building out new missile warning and tracking satellite systems and a long-range discriminating radar to improve the efficacy of the existing missile defense system. Additional missile defense capabilities like fielding more interceptors, interceptors with multiple kill vehicles, or directed energy capabilities could improve U.S. mean damage limitation further, although these systems would also likely increase uncertainty over U.S. damage limitation.

In Figure 13 we evaluate the utility of changing the variance values for the U.S., versus changing variance around Chinese damage limitation capability. This is to help understand which measure is more valuable for the U.S. Here both contour plots show a much more significant benefit for small increases in variance around U.S. damage limitation, suggesting that Washington can gain more utility in creating uncertainty for China about what specifically the U.S. can or cannot do, than it can by gaining more insight on China's true capabilities. We see that very small increases in uncertainty around the U.S. capability are equivalent in utility to

much larger decreases in uncertainty about China's capabilities, roughly 2-3 times more depending on the symmetry of stakes of the crisis. This is likely because of the slightly higher U.S. mean damage limitation values (in this case, 0.5 for the U.S. versus 0.3 for China), resulting in equal changes in uncertainty being able to drive U.S. utility higher.

In this context, marginal improvements in strategic missile defenses would likely improve the U.S. bargaining position even if they only resulted in additional Chinese uncertainty about the U.S. capability. The U.S. has already declared it is pursuing "left of launch" capabilities, to include cyber or electronic warfare capabilities, to help bolster and expand missile defenses.[44] These capabilities would also likely simply increase China's uncertainty about U.S. damage limitation, but revealing the expansion of these capabilities would likely also improve the U.S.' utility and its chances of prevailing in a crisis.

Figure 12. Values for U.S. expected utility (with lighter colors showing higher values) in trading U.S. raid sizes (y-axis) against changes in mean damage limitation (x-axis). Left table shows this under symmetric stakes crises, while right graph shows this under asymmetric stakes (China advantage). Expected utility takes into account both probability of winning and probability of ruin. Values are given for the 2036 expected force balance with China under the assumption of a high stakes crisis, and a perfectly effective Chinese launch under attack capability.

[44] U.S. Department of Defense 2018.

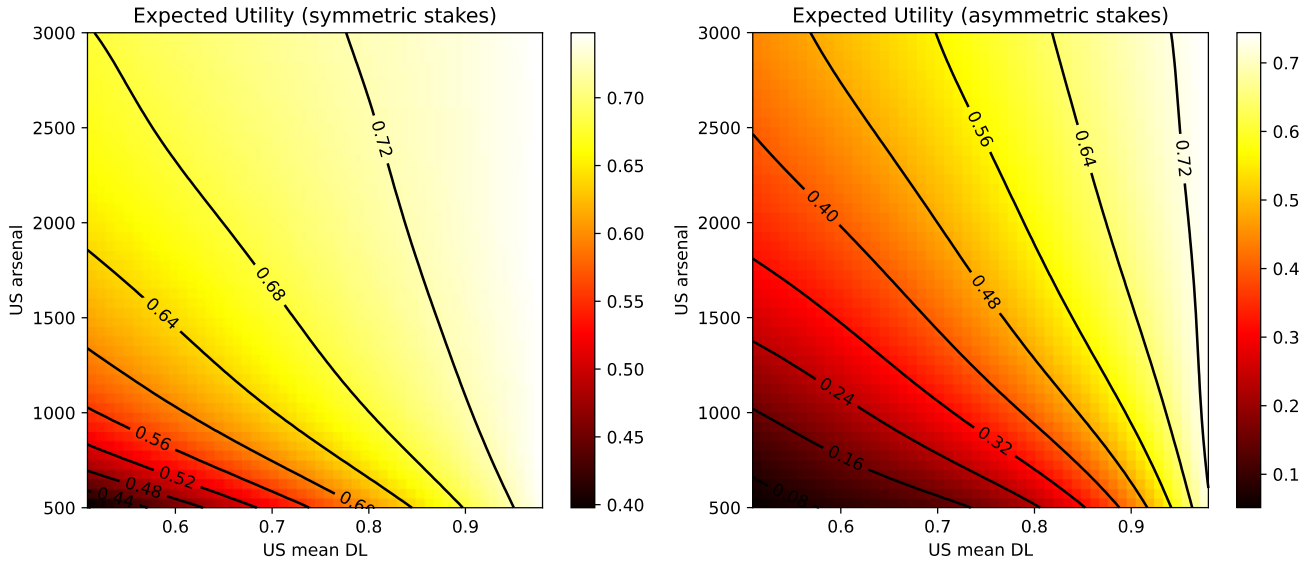
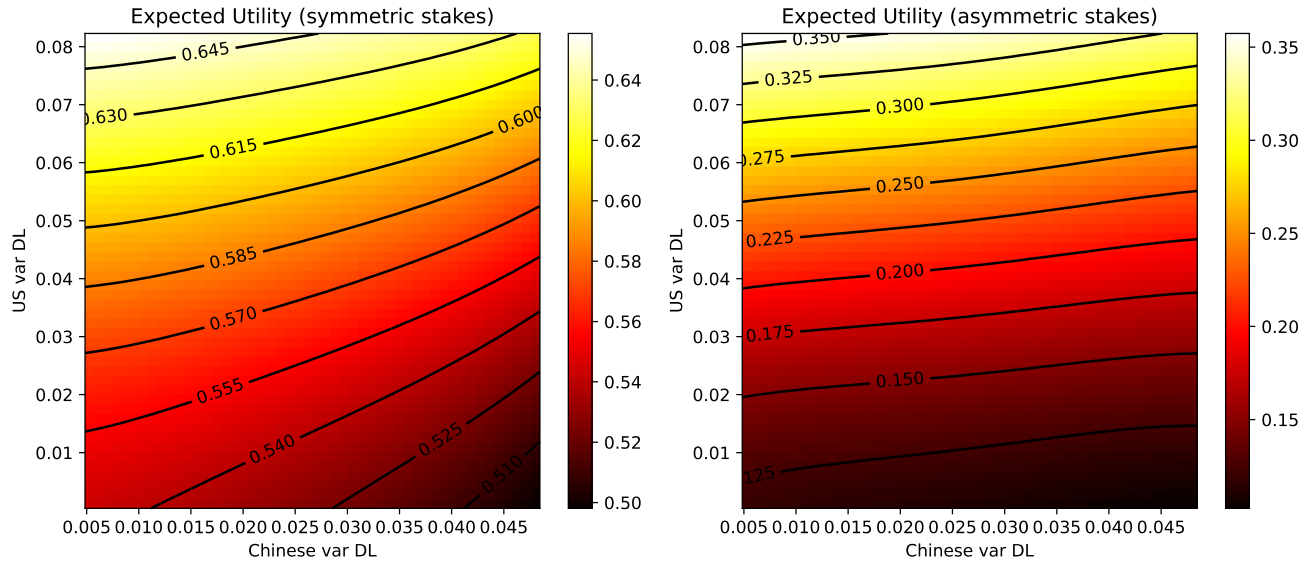


Figure 13. Values for U.S. expected utility when changing uncertainty around U.S. damage limitation (y-axis) versus changes in uncertainty (x-axis) about Chinese damage limitation. All formats and assumptions same as Figure 12.



Conclusion

The preceding analysis shows that even modestly effective damage limitation capabilities can impact crisis bargaining, even when there is a simple advantage in uncertainty about the

balance of power. We show that our model captures different dynamics that Powell's model, and that uncertainty around damage limitation capabilities can drive outcomes differently than uncertainty around the players' resolve. Our analysis elucidates the fundamental logic of the delicate balance theory of deterrence under conditions of significant uncertainty by using a formal game theory model which allow us to assess the impact of new or additional capabilities on bargaining outcomes. In our model, the combined set of missile defense and counterforce capabilities create an asymmetry in the perceived balance of survivable strategic forces, which thus acts as an expected cost differential that the more capable side can bargain with. The asymmetric decrease in expected cost of war that the stronger player faces therefore increases that player's effective resolve, which allows them to bid higher in brinksmanship bargaining. But this asymmetry of risk also simultaneously cows the weaker player, decreasing their bids, as they fear that their arsenal will have significantly diminished viability if the crisis escalates to war.

The dynamics we observe in our model results could explain the more restrained behavior of the Soviets when they perceived that U.S. had the ability to significantly attrit their nuclear forces, which Richard Betts elucidates in his seminal work on crisis bargaining cases in the nuclear era.[45] And further, it could also help explain the Soviet leadership's aggressive moves to develop more survivable nuclear forces in the 1970s and 1980s as these forces would have improved their crisis bargaining position.[46] Our findings also provide some mathematical validation of early ideas about strategic advantage and crisis bargaining posited by Wohlstetter and Kahn and more recently espoused by Green, Long and Talmadge.

[45] Betts, 1999.
[46] Green and Long, 2017b.

In contrast to much of the scholarly literature to date, our findings suggest that damage limitation capabilities could be a force for crisis stability as they induce challengers to drop out of crises earlier because of their weakened bargaining position. Our analysis of the probability of a general nuclear exchange suggests that the chances of mutual annihilation are low under conditions of the delicate balance, especially when damage limitation capabilities are highly asymmetric between players. We find that this is particularly true in cases where the political stakes are high, as asymmetric damage limitation capability shifts the probability of a general nuclear exchange downward by making the stronger party bid higher and the weaker party bid lower. After all, in the traditional game of chicken under nuclear stalemate there is significant risk of both parties standing firm if the stakes are high, so the probability of a general nuclear exchange is also relatively high. Our analysis shows that the impact of these capabilities on the risk of mutual destruction is via both emboldening the stronger side as well as cowing of the weaker side—which cannot run these same risks—making them far more likely to submit in equilibrium.

Our findings have important implications for future empirical and theoretical research. More empirical work is needed to understand bargaining behavior in past nuclear crises. Our analysis suggests that more evidence is needed on Soviet decision making in crises to understand their calculus, particularly to validate the idea that they were cowed due to their disadvantage in the nuclear balance of power. Campaign analysis techniques also need to be extended to handle nuclear weapons exchanges using air and sea-based weapons, which the Riqiang and Tecott-Metz Halterman models assume the U.S. can perfectly attrit, as these dynamics will become more salient as China's sea and air-based nuclear forces grow and mature in the 2020s.

Scholars should also examine these bargaining dynamics with theoretical models

incorporating sequential moves in a crisis, as well as models where players have repeated interactions. Our findings suggest that the weaker party will tend to avoid serious crises in the face of a weak bargaining position, as others have also found in research on crisis entry conditions with sequential game models.[47] However, this needs further study as we did not evaluate the complexities of crisis entry conditions in this analysis. Our model cannot capture sequential move game dynamics like runaway escalation cycles, tit-for-tat strategies, or "use or lose" incentives. But one specific advantage of our type of model is that the uncertainty inherent in it could help drive the players into crisis in the first place, especially compared to other models that assume perfect information on both sides about the balance of power. Additional theoretical research to specifically analyze crisis entry conditions would better elaborate the implications for general deterrence.

These findings also have important policy implications. First, our results show that the U.S. government has a number of levers it can pull to adjust its nuclear posture to help reduce China's future coercive leverage in a crisis. Our results show that the United States would benefit more from modest improvements in its damage limitation capability, in comparison to even large increases in deployed nuclear weapons or upload capability. Washington can achieve this by increasing warhead accuracy to drive up single shot kill probability, by improving mid-course missile discrimination and thereby improving missile defense efficacy, or by successfully upgrading the national missile defense system with new intercept capabilities. Keeping China's damage limitation capability below 30% is also important, meaning that investments in the survivability of U.S. strategic missile submarines are critical, as is maintaining the ability to upload multiple warheads on U.S. bombers and potentially ICBMs (vs. the single warhead they

[47] Wagner 1991.

currently carry)[48] in case China makes breakthroughs in anti-submarine warfare.

But our analysis also shows that there are significant benefits to Washington manipulating uncertainty around its own damage limitation capability, and to a lesser degree in developing additional intelligence to reduce its uncertainty about future Chinese damage limitation. While we cannot judge the value of the marginal dollar of investment in our model, our findings suggest that perhaps a weighted combination of these measures is appropriate to create competitive advantages across the 2036 time epoch and beyond.

The long-run risks to pursuing an advantage in the delicate balance framework are twofold. First, preventing an arms race is the key goal for U.S.-China competition. Arms race dynamics will likely be a growing problem both as China's arsenal grows and as the U.S.-Russia New START treaty is expected to expire in 2026. One risk in our model of competition is that growth of one side's counterforce and missile defense capabilities incentivizes the other side to field greater numbers of weapons and delivery systems, reducing the ability of stronger player's mean effectiveness—and potentially triggering an arms race spiral. Countries like China could find themselves under increasing pressure to adopt a more coercive nuclear strategy to face down this shift.[49] However, our model shows a number of avenues around quantitative competition. First, both sides could increase defensive investments via hardening and survivability of nuclear weapon systems and command-and-control to buy back advantage. They can also pursue other advantages like missile defense and weapon accuracy, which are qualitative in nature. Third, competition in the areas where uncertainty plays a key role, like in missile defenses, weapon accuracy, space, and non-kinetic weapons like cyber, may be less prone to arms racing even if it

[48] U.S. Navy, 2023.
[49] Lieber and Press 2020, p. 108-110, 117-118.

threatens hardened nuclear platforms like ICBMs. This is because it is less clear how to invest to make these systems hardened or more resilient against such amorphous threats. And last, our model also shows that there will be significant benefits to the side with superior intelligence capabilities, which similarly avoids direct arms race spiral dynamics.

The second and more long-run risk is that both competitors develop significant missile defense and counterforce capabilities in a damage limitation arms race spiral. Our model shows that if both sides have damage limitation capabilities that are more than 30% effective, the risks of ruin begin to increase significantly; and as these capabilities increase to roughly 70-80% effectiveness, the game becomes more akin to a prisoner's dilemma as the risk of mutual annihilation grows significantly. One mitigation to this risk is for states to first prioritize defensive investments to secure their own nuclear forces and command and control. However, we should note that highly effective Chinese counterforce and missile defense capabilities are unlikely to emerge before 2036, so this risk can be mitigated in the 2030s.

Last, there are several implications that flow from this analysis regarding extended deterrence and first-strike stability. Our findings show that Russian and Chinese complaints that U.S. missile defense and counterforce capabilities are a threat to strategic stability miss the mark, as asymmetric U.S. capabilities clearly lower the risk of mutual destruction. Our findings suggest that how China and other regional challengers perceive the U.S.'s damage limiting capabilities could be key to the U.S.'s success in extended deterrence via reducing the challengers' perception of winning if the crisis escalates significantly. Our analysis also suggests that both the balance of power and the balance of stakes are key determinants in crisis bargaining outcomes, so managing perceptions of U.S. stake is critical for maintaining general deterrence and crisis stability. Efforts to highlight the importance of U.S. treaty commitments will

demonstrate that the U.S. has a significant stake in any crisis involving a treaty ally or close partner, as the failure of any one U.S. commitment would endanger the integrity of all, and this would help maintain the perception that any future crisis would be relatively symmetric in stakes.

REFERENCES

James M. Acton, "Escalation through Entanglement." *International Security*, Vol. 43, No. 1 (Summer 2018).

America's Strategic Posture: The Final Report of the Congressional Commission on the Strategic Posture of the United States, (Washington DC; Institute for Defense Analyses, 2023).

Richard Betts, *Nuclear Blackmail and Nuclear Balance*, (Brookings Institution Press, 1987).

Stephen J. Cimbala, "Nuclear Crisis Management and 'Cyberwar': Phishing for Trouble?", *Strategic Studies Quarterly,* Spring 2011, p. 117-128.

Christopher Clary, "Survivability in the New Era of Counterforce," in Narang and Sagan, eds, *The Fragile Balance of Terror: Deterrence in the New Nuclear Age,* (Cornell, N.Y., Cornell University Press, 2022).

Fiona S. Cunningham, review in H-Diplo Roundtable XXIII-11, 12 November 2021.

Fiona S. Cunningham and M. Taylor Fravel, "Assuring Assured Retaliation: China's nuclear posture and U.S.-China strategic stability," *International Security*, Vol. 40, No. 2 (Fall 2015), pp. 7–50.

Abby Fanlo, Lauren Sukin, "The Disadvantage of Nuclear Superiority," *Security Studies*, 2023.

Erik Gartzke and Jon R. Lindsay, editors, *Cross Domain Deterrence: Strategy in an Era of Complexity* (Oxford, U.K.; Oxford University Press, 2017).

Erik Gartzke and Jon R. Lindsay, "Thermonuclear cyberwar," *Journal of Cybersecurity*, 3:1, (March 2017), pp. 37–48.

Francis J. Gavin, "Strategies of Inhibition: U.S. Grand Strategy, the Nuclear Revolution, and Nonproliferation," *International Security*, Vol. 40, No. 1, (Summer 2017).

Charles L. Glaser and Steve Fetter, "Counterforce Revisited: Assessing the Nuclear Posture Review's New Missions," *International Security,* Vol. 30, No. 2, (Fall 2005).

Charles Glaser and Steven Fetter, "Should the United States Reject MAD? Damage Limitation and U.S. Nuclear Strategy toward China," *International Security,* Vol. 41, No. 1, (Summer 2016)

Charles Glaser, James Acton, Steve Fetter, "The U.S. Nuclear Arsenal can Deter Both Russia and China: Why America Doesn't Need More Missiles," *Foreign Affairs* magazine, October 2023.

Brendan Rittenhouse Green and Austin Long, "Conceal or Reveal? Managing Clandestine Military Capabilities in Peacetime Competition," *International Security*, Vol. 44, No. 3, (2020).

Brendan Rittenhouse Green and Austin Long, "Correspondence: The Limits of Damage Limitation," *International Security*, Vol. 42, No. 1, (Summer 2017).

Brendan Rittenhouse Green and Austin Long, "The MAD Who Wasn't There: Soviet Reactions to the Late Cold War Nuclear Balance," *Security Studies*, Vol 26, No. 4 (2017).

Brendan Rittenhouse Green, *The Revolution that Failed: Nuclear Competition, Arms Control and the Cold War,* (Cambridge, UK; Cambridge University, 2020).

Brendan Rittenhouse Green, Austin Long, Mark S. Bell, Julia Macdonald, "Contrasting Views on How to Code a Nuclear Crisis," *Texas National Security Review*, Vol 2, No. 4, October 2019, 130–139.

Ken Hendricks, Andrew Weiss, and Charles Wilson, "The War of Attrition in Continuous Time with Complete Information," *International Economic Review*, Vol. 29, No. 4 (November 1988).

Heinrich Stalhane Hiim, M. Taylor Fravel, Magnus Langset Troan, "The Dynamics of an Entangled Security Dilemma: China's Changing Nuclear Posture," *International Security*, Vol 47, No. 4, (Spring 2023).

Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Cornell, NY; Cornell University Press, 1989).

Herman Kahn, *On Escalation: Metaphors and Scenarios*, (Transaction Press, 1965).

Matthew Kroenig, "Nuclear Superiority and the Balance of Resolve: Explaining Nuclear Crisis Outcomes," *International Organization*, Vol. 67, No. 1 (Winter 2013).

Matthew Kroenig, *the Logic of American Nuclear Strategy*, (Oxford, UK; Oxford University Press, 2018).

Dana Higgins, Connor Huff, Anton Strezhnev, "Survivability not Superiority: A Critique of Kroenig (2013)" Working Paper.

Keir Lieber and Daryl Press, *US Strategy and Force Posture for an Era of Nuclear Tripolarity*, (Washington DC; Atlantic Council, 2023).

Keir Lieber and Daryl Press, *The Myth of the Nuclear Revolution: Power Politics in the Atomic Age*, (Ithaca, N.Y.: Cornell University Press, 2020).

Keir Lieber and Daryl Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security*, Vol. 41, No. 4, (2017).

Study Group Convened by Lawrence Livermore National Laboratory, *China's Emergence as a Second Nuclear Peer: Implications for U.S. Nuclear Deterrence Strategy,* (Livermore, C.A.: Lawrence Livermore National Laboratory Center for Global Security Research, 2023).

Barry Nalebuff, "Brinkmanship and Nuclear Deterrence: The Neutrality of Escalation," *Conflict Management and Peace Science*, Vol. 9, No. 2, (1986).

Robert Powell, "Nuclear Deterrence Theory, Nuclear Proliferation, and National Missile Defense," *International Security*, Vol. 27, No. 4 (Spring 2003).

Wu Riqiang, "Living with Uncertainty: Modeling China's Nuclear Survivability," *International Security*, Vol. 44, No. 4, (Spring 2020).

Brad Roberts, *On Theories of Victory: Red and Blue* (Livermore, CA; Lawrence Livermore National Laboratory, 2020).

Rumsfeld Commission report, "Report of the Commission to Assess the Ballistic Missile Threat to the United States," (Washington D.C.: U.S. Congress, 1998). Donald H. Rumsfeld, "Prepared Testimony to the Senate Armed Services Committee," June 21, 2001, p.2, http://www.comw.org/qdr/010621rumsfeld.pdf

Thomas Schelling, *The Strategy of Conflict*, (Cambridge, MA; Harvard University, 1960).

Peter Schram, "Hassling: How States Prevent a Preventive War," *American Journal of Political Science*, Vol. 65, No. 2 (April 2021), pp. 294-308.

Todd S. Sechser, Review of "The Logic of American Nuclear Strategy: Why Strategic Superiority Matters" By Matthew Kroenig. New York: Oxford University Press, 2018, *Perspectives on Politics*, Vol 17, No 4, November 2019.

Todd S. Sechser and Matthew Fuhrmann, "Crisis Bargaining and Nuclear Blackmail," *International Organization*, Vol. 67, No. 1, (January 2013), pp. 173 – 195.

Braden Soper, "A Cyber-Nuclear Deterrence Game." *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, 2019, pp. 470-479.

Caitlin Talmadge, "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States," *International Security*, Vol. 41, No. 4 (Spring 2017), pp. 50–92.

Caitlin Talmadge, "Multipolar Deterrence in the Emerging Nuclear Era," in Narang and Sagan, eds, *The Fragile Balance of Terror: Deterrence in the New Nuclear Age*, (Cornell, N.Y., Cornell University Press, 2022).

Rachel Tecott Metz and Andrew Halterman, "The Case for Campaign Analysis: A Method for Studying Military Operations," *International Security*, Vol. 45, No. 4, (Spring 2021).

U.S. Department of Defense, "Declaratory Policy, Concept of Operations and Employment Guidelines for Left-of-Launch Capability," May 10, 2017.

U.S. Department of Defense, Office of the Secretary of Defense, "Annual Report to Congress: Military and Security Developments Involving the People's Republic of China," 2023.

U.S Department of Defense, Missile Defense Review, 2019.

U.S Department of Defense, Missile Defense Review, 2022.

U.S. Navy Website, https://www.sublant.usff.navy.mil/About-Us/Submarine-Facts/, website accessed July 6, 2023.

Harrison Wagner, “Nuclear Deterrence, Counterforce Strategies, and the Incentive to Strike First,” American Political Science Review, Vol. 85, No. 3, (1991).

Albert Wohlstetter, “The Delicate Nuclear Balance,” *Foreign Affairs* Magazine, 1959.

Technical Appendix

In this appendix we provide more technical details on various aspects of the game theory model developed in the paper. Here we first provide more detailed background on the second-price, all-pay auction models of Nalebuff and Powell, giving the first-order equilibrium conditions of the game and the general form of equilibrium solutions. To extend this model to include generalized damage limitation effectiveness, we introduce specific distributional assumptions on both baseline resolve and damage limitation effectiveness, leading to our distributions over a player's effective resolve. Finally, we end with a discussion on how to incorporate more detailed campaign analysis, including specific measures such as nuclear arsenal size and reliability, into our modeling framework.

ADDITIONAL DETAIL ON OUR MODEL DERIVATION

Denote the strategic risk for player i by $r_i \in [0,1]$. Given a strategy profile $(r_1, r_2) \in [0,1] \times [0,1]$ we can determine the expected payoffs for both players. For example, if $r_1 > r_2$ then the expected payoff for player 1 is $P_1 = w_1(1 - r_2) - d_1 r_2$ while the expected payoff for player 2 is $P_2 = -s_1(1 - r_2) - d_2 r_2$. When $r_1 < r_2$ we have $P_1 = -s_1(1 - r_2) - d_1 r_2$ and $P_2 = w_1(1 - r_2) - d_2 r_2$. If $r_1 = r_2$ we assume $P_1 = -s_1(1 - r_2) - d_1 r_2$ and $P_2 = -s_1(1 - r_2) - d_2 r_2$.

Recall that in both the Nalebuff and Powell second-price, all-pay auction models, strategies are determined by what Powell terms the player's resolve, $R_i = \frac{w_i}{w_i + d_i}$. In this way, all private information can be summarized by the player's resolve, and specifying the game requires specifying distributions over the players' resolve. Let $F_i(x) = P(R_i < x)$ denote the cumulative distribution function of the probability distribution on player i's type (resolve). Note that this

represents player $-i$'s beliefs about player i's resolve. We denote the density function by $f_i(x) = \frac{d}{dx}F_i(x)$. Given player beliefs about opponent resolve (represented by distributions F_i), we can write down the expected payoff for player i with resolve R_i when choosing a risk level r. To do so we will need to define the inverse functions $r_i^{-1}(r)$ as the level of player i resolve which chooses to play a level of risk r. For example, suppose player i has a resolve of R_i and plays strategy r_i: $[0,1] \to [0,1]$, and according to this strategy $r_i(R_i) = x$ for some value $x \in [0,1]$. Then we define the inverse function r_i^{-1}: $(0,1) \to [0,1]$ as the function satisfying $r_i^{-1}(x) = R_i$.[50] With this notation we can write the expected payoff for player i with resolve R_i when choosing a risk level r as follows.

$$V_i(r \mid R_i) = -(s_i(1-r) + d_i r)\,(1 - F_{-i}(r_{-i}^{-1}(r))) + \int_0^{r_{-i}^{-1}(r)} [w_i\big(1 - r_{-i}(u)\big) - d_i r_{-i}(u)] f_{-i}(u)du$$

The first term is the expected payoff to conceding which takes into account the probability that the opponent chooses a risk level greater than r (according to player i's beliefs). The second term is the expected payoff to prevailing which takes into account the probability that the opponent chooses a risk level less than r (according to player i's beliefs).

Differentiating the expected payoffs $V_i(r \mid R_i)$ with respect to risk strategy r and solving $\frac{d}{dr}V_i(r \mid R_i)$=0 for r for $i = 1,2$ yields first-order conditions for an equilibrium profile (r_1^*, r_2^*). Nalebuff shows that solving the first-order conditions reduces to solving the following ordinary differential equation which defines R_i as a function of R_{-i}.

[50] Note that the inverse function $r_i^{-1}(r)$ is always well defined on the open interval $(0,1)$ based on the continuity and monotonicity of the function $r_i(R)$.

$$\frac{1-R_i}{R_i}\frac{f_i(R_i)}{1-F_i(R_i)}\frac{dR_i}{dR_{-i}} = \frac{1-R_{-i}}{R_{-i}}\frac{f_{-i}(R_{-i})}{1-F_{-i}(R_{-i})}$$

Let $\hat{R}_i(R_{-i})$ be a solution to this ODE. Then $\hat{R}_i(R_{-i})$ is the level of resolve that player i must have been to have chosen the same level of risk (in equilibrium) that player $-i$ chose (in equilibrium) when her resolve was R_{-i}. Note that solving the ODE for R_{-i} as a function of R_i leads to the analogous function $\hat{R}_{-i}(R_i)$. The general form of the equilibrium solution to these equations can then be written as follows.

$$r_i^*(R_i) = 1 - \exp\left\{-\int_0^{\hat{R}_{-i}(R_i)} \frac{\hat{R}_i(R_{-i})}{1-\hat{R}_i(R_{-i})}\frac{f_{-i}(R_{-i})}{1-F_{-i}(R_{-i})}dR_{-i}\right\} \quad (1)$$

Note that closed form analytical solutions are only available for a very limited class of functions F_i.

Powell builds on the above model in two ways. The first, and most relevant extension for our purposes, is to assume that one player (the U.S., in Powell's game) possesses a national missile defense system capable of stopping a nuclear attack with probability of success $e \in [0,1]$. If player i has the national missile defense capability, then the cost of a nuclear attack is reduced to the expected cost $(1-e)d_i$. Powell further extends this model to include preliminary actions made by both the U.S. and a nuclear armed rogue state. It is assumed the rogue state makes some initial move to initiate a crisis. The U.S. must decide whether to engage the rogue state by also entering into the crisis, while the rogue state must decide whether to remain in the crisis if the U.S. engages. If the U.S. engages and the rogue state remains, then a nuclear brinkmanship auction game similar to the one presented above unfolds. Power derives the expected payoffs and first-order equilibrium conditions for this extended brinkmanship model. Assuming player i has

the national missile defense capability with effectiveness e, the equivalent first-order conditions can be reduced to the following ordinary differential equation.

$$(1-e)\frac{1-R_i}{R_i}\frac{f_i(R_i)}{1-F_i(R_i)}\frac{dR_i}{dR_{-i}} = \frac{1-R_{-i}}{R_{-i}}\frac{f_{-i}(R_{-i})}{1-F_{-i}(R_{-i})}$$

As discussed in the main body of the paper, we use effective resolve R_i^* as the player type. This is justified by the fact that all equilibria can be characterized by the effective resolve R_i^*. This follows directly from the analysis in Nalebuff by simply replacing d_i with $(1-e_i)d_i$. Assuming distributions F_i over R_i^* the relevant ordinary differential equation is as follows.

$$\frac{1-R_i^*}{R_i^*}\frac{f_i(R_i^*)}{1-F_i(R_i^*)}\frac{dR_i^*}{dR_{-i}^*} = \frac{1-R_{-i}^*}{R_{-i}^*}\frac{f_{-i}(R_{-i}^*)}{1-F_{-i}(R_{-i}^*)} \quad (2)$$

To fully specify our model, we need to specify probability distributions over effective resolve. Because both baseline resolve and damage limitation effectiveness have support on the unit interval, we choose beta distributions to model player beliefs around these values. We write Beta(a,b) to denote a beta distribution with support [0,1] and shape parameters $a>0, b>0$. Then given shape parameters $\alpha_i>0$, $\beta_i>0, \gamma_i>0$, $\lambda_i>0$ for $i=1,2$, we can define a distribution over effective resolve R_i^* as follows. Let the baseline resolve of player i be beta distributed, $R_i \sim$ Beta(α_i, β_i), and let the damage limitation effectiveness of player i be beta distributed, $e_i \sim$ Beta(γ_i, λ_i). We then assume that the beliefs about baseline resolve R_i and effectiveness e_i are independent. We can derive the distribution of effective resolve using the relation $R_i^* = \frac{R_i}{R_i+(1-e)(1-R_i)}$. Denoting the probability density function of effective resolve for player i by $f_i(R_i^*)$ we have the following.

$$f_i(R_i^*) = \int_0^1 \frac{R_i^{*\alpha_i-1}(1-R_i^*)^{\beta_i-1}}{B(\alpha_i,\beta_i)} \frac{e_i^{\gamma_i}}{(1-e_iR_i^*)^{\gamma_i+\lambda_i}} \frac{e_i^{\gamma_i-1}(1-e_i)^{\lambda_i-1}}{B(\gamma_i,\lambda_i)} de_i \quad (3)$$

For a beta distributed random variable $x \sim \text{Beta}(a,b)$ with shape parameters a and b, the mean of x is given by $E[x] = \frac{a}{a+b}$ and the variance of x is given by $Var[x] = \frac{ab}{(a+b)^2(a+b-1)}$. Thus, the ex ante mean and variance of player i's baseline resolve are $E[R_i] = \frac{\alpha_i}{\alpha_i+\beta_i}$ and $Var[R_i] = \frac{\alpha_i\beta_i}{(\alpha_i+\beta_i)^2(\alpha_i+\beta_i-1)}$, respectively. Similarly, the ex ante mean and variance of player i's damage limitation effectiveness are $E[e_i] = \frac{\gamma_i}{\gamma_i+\lambda_i}$ and $Var[e_i] = \frac{\gamma_i\lambda_i}{(\gamma_i+\lambda_i)^2(\gamma_i+\lambda_i-1)}$, respectively. With this we can now explore what happens when changes occur to either the mean or the variance of both baseline resolve and damage limitation effectiveness.

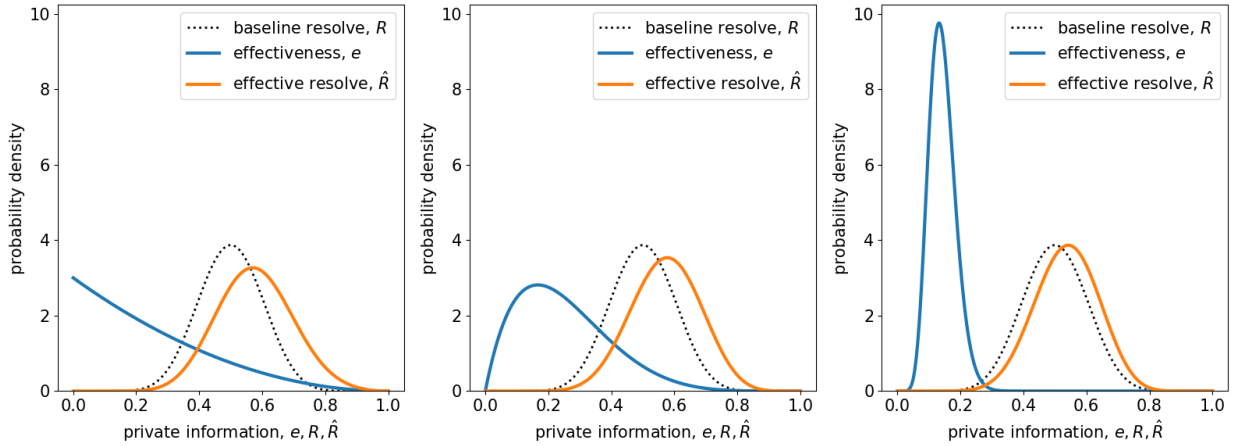


Figure 1. Three examples distributions of effective resolve along with the associated distributions of baseline resolve and damage limitation effectiveness. In each subfigure the baseline resolve and mean damage limitation is fixed, which the variance of the damage limitation is increased from left to right. It is clear that high variance shifts effective resolve towards the right (higher effective resolve).

NUMERICAL IMPLIMENTATION

The distribution on effective resolve given in equation (3) is not available in closed form and must be approximated numerically. This means that equilibrium solutions given by (1) must be approximated numerically as well. Because we are interested in performing a parameter study, this requires solving (1) at many different points in the eight dimensional parameter space of the game. To do this efficiently we leveraged high performance computing to solve for many game equilibria in parallel.

We first provide some details on the numerical approximation of the game equilibria given in (1) when player type is provided by the distributions in (3) for some fixed set of game parameters, $(\alpha_1, \beta_1, \gamma_1, \lambda_1, \alpha_2, \beta_2, \gamma_2, \lambda_2)$. We first note that approximating the integral in (1) directly using the definition of (3) is not numerically stable. For this reason we first used a kernel density estimation of (3) using samples directly sampled from beta distributions. In particular, we sampled values $e_{ij} \sim \text{Beta}(\gamma_i, \lambda_i)$ and $R_{ij} \sim \text{Beta}(\alpha_i, \beta_i)$ for $i = 1,2$ and $j = 1,2, \dots, 1000$. We then compute $R^*_{ij} = \frac{R_{ij}}{R_{ij}+(1-e_{ij})(1-R_{ij})}$ to get samples from the distribution implied by the pdf in (3). Using the samples R^*_{ij} we used a Gaussian kernel density estimator for (3). This approximation proved much more reliable for approximating the cdf $F_i(R^*_i)$ and subsequently the equilibrium solutions in (1). The cdf was approximated as follows. First we used Gaussian quadtrature to approximate $v_k = \int_0^{x_k} f_i(u)du$ where the values $x_k = k/1000$. Next we used linear interpolation on the points (x_k, v_k) to approximate the cdf $F_i(R^*_i)$.

The next step is to approximate the function $\hat{R}^*{}_i(R^*_{-i})$, which is the function that solves the ODE in (2). Recall this function maps player $-i$'s type to player i's type which has the same equilibrium bidding strategy, i.e. $r_i(\hat{R}^*{}_i(R^*_{-i})) = r_{-i}(R^*_{-i})$. Following Nalebuff the function $\hat{R}^*{}_i(R^*_{-i})$ is defined by $\hat{R}^*{}_i(R^*_{-i}) = H_i^{-1}(H_{-i}(R^*_{-i}) + k_{-i})$ where k_{-i} is a constant of integration

and $H_i(R_i^*) = \int_{\beta}^{R_i^*} \frac{(1-y)f_i(y)}{y(1-F_i(y))} dy$ for some arbitrary constant $\beta \in [0,1]$. The function H_i was approximated with Gaussian quadrature and H_i^{-1} was approximated with the bisect method and the approximation of H_i, i.e., if $H_i^{-1}(y)$ was needed we used the bisect method to find the root x_0

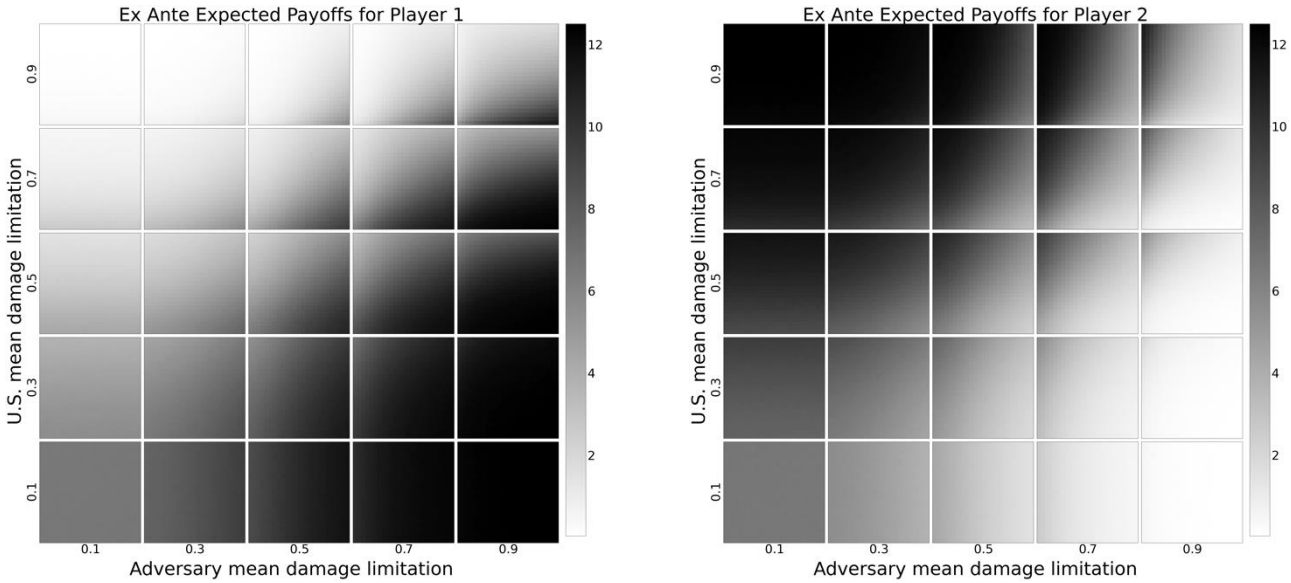


Figure 2. Ex ante equilibrium costs for players 1 (left, nominally the U.S.) and 2 (right, nominally an adversary). Baseline resolve for both players is fixed at Beta(1,9). Damage limitation varies along x and y axis. In each subplot, the mean damage limitation is fixed at the denoted level. The variance of damage limitation is increased along the x axis for player 2 and along the y axis for player 1.

of the function $g_y(x) = H_i(x) - y$. Using these approximations to $H_i(x)$ and $H_i^{-1}(x)$ for $i = 1,2$ it was possible to approximate the functions $\hat{R}^*{}_i(R^*_{-i})$ for $i = 1,2$. With all of these functions approximated we could then approximate (1) using the distributions in (3).

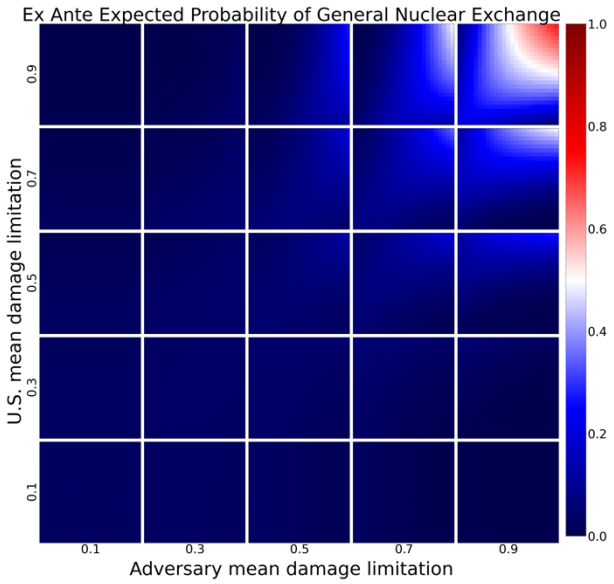


Figure 3. Ex ante equilibrium expected probability of ruin for player 1 (nominally the U.S.). Baseline resolve for both players is fixed at Beta(1,9). Damage limitation varies along x and y axis. In each subplot, the mean damage limitation is fixed at the denoted level. The variance of damage limitation is increased along the x axis for player 2 and along the y axis for player 1.

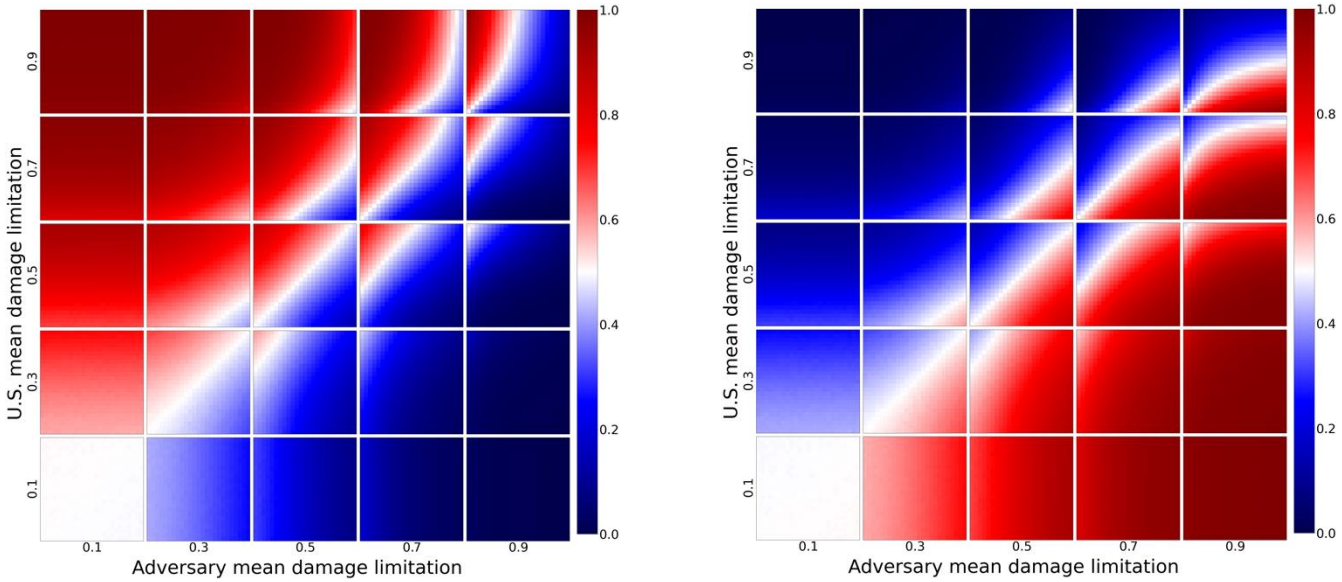


Figure 4. Ex ante equilibrium expected probability of prevailing for player 1 (left, nominally the U.S.) and 2 (right, nominally an adversary). Baseline resolve for both players is fixed at Beta(1,9). Damage limitation varies along x and y axis. In each subplot, the mean damage limitation is fixed at the denoted level. The variance of damage limitation is increased along the x axis for player 2 and along the y axis for player 1.

CAMPAIGN ANALYSIS

In this section we show how we use Wu Riqiang's campaign analysis model to determine parameter values for our game theory model.[51] Let N_c be the total number of deployed Chinese nuclear weapons and let N_u be the total number of deployed U.S. nuclear weapons. In all years considered in this analysis we have $N_c < N_u$. Denote the number of deployed Chinese nuclear weapons that survive a U.S. first strike by N_c^s. In Riqiang's analysis a Monte Carlo (MC) simulation of the campaign analysis model produces an estimate of the probability distribution over the number of deployed Chinese nuclear weapons that survive a U.S. first strike, i.e., the MC simulations produce samples of N_c^s.

We relate the number of surviving Chinese warheads N_c^s to U.S. damage limitation e_u as follows. Assume that the U.S. assigns a value v_u to each potential location the Chinese are targeting. If the U.S. had no damage limitation, then the maximum damage it could incur is $v_u N_c$. If the U.S. had damage limitation $e_u > 0$, then the damage it could incur is $(1 - e_u)v_u N_c$. In the Reqiang model the total realized damage to the U.S. is $v_u N_c^s$. Thus equating total realized damage in the two models allows us to relate the Reqiang model results to our models parameters. Namely, we set

$$(1 - e_u)v_u N_c = v_u N_c^s.$$

Solving for damage limitation gives us the following.

$$e_u = 1 - \frac{N_c^s}{N_c}$$

[51] Wu Riqiang, "Living with Uncertainty: Modeling China's Nuclear Survivability," *International Security*, Vol. 44, No. 4, (Spring 2020)

We now see that a distribution on N_c^s implies a distribution on e_u. Wu gives estimates of the *survival function* of N_c^s, which implies a distribution over e_c when N_c is known. To approximate this distribution in the context of our mode we look for beta parameters a, b that best match the distribution implied by Wu.

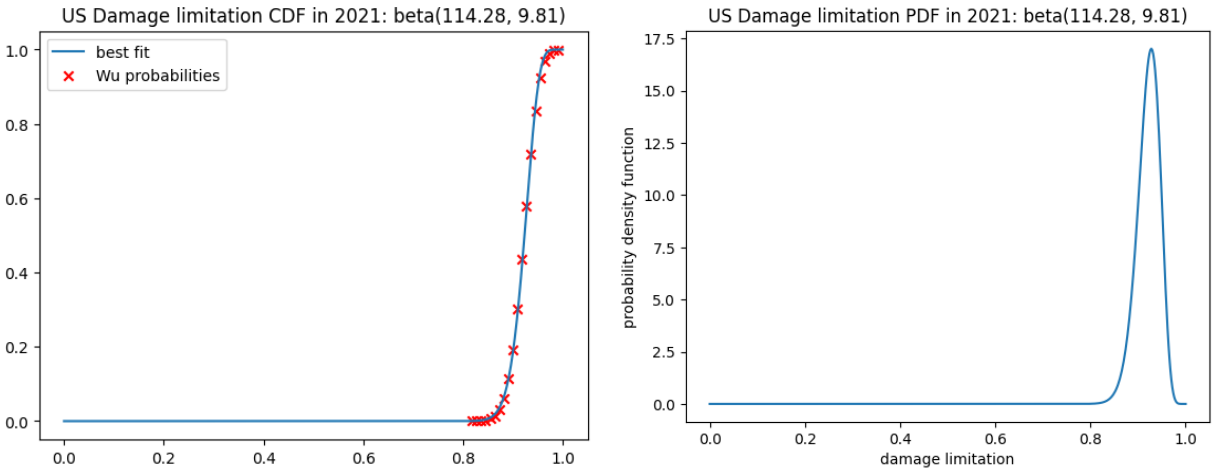


Figure 5. The left figure is the cumulative distribution function (CDF) of the U.S. damage limitation. The red marks are the probabilities that come out of the MC simulations of the Reqiang model. Note that we have scaled the x axis so that it is on the damage limitation scale. The blue line is the CDF of the best fit beta distribution. The right figure shows the corresponding probability distribution function (PDF).

Let $S_c(e)$ be the survival function over damage limitation implied by Wu's analysis. Let $S_\beta(e, a, b)$ be the survival function of a beta distribution with parameters a, b. Then we want to minimize $\sum_i \left(S_c(e_i) - S_\beta(e_i, a, b)\right)^2$ with respect to the parameters a, b. Thus for any assumption on the number of nuclear weapons, the above analysis allows us to specify the U.S. damage limitation. In order to specify the U.S. baseline resolve we recall that $R = \frac{w}{w+d}$. Assume the stakes $w \sim \text{Gamma}(a_u, b_u)$ for some parameters a_u, b_u. Furthermore, as China increases the number of deployed weapons, the total damage d should increase. To this end we set $d = N_c$. The implicit assumption is that any subjective costs associated with the damage incurred by the N_c weapons have been absorbed into the uncertainty around w. We make analogous assumptions

around the baseline resolve of China. Namely, $d_c = N_u$ and $w \sim \text{Gamma}(a_c, b_c)$ for some parameters a_c, b_c. This define probability distributions over baseline resolve R_u and R_c. In order to relate these distributions to our model, we again find the best matching beta distribution. To do this we draw i.i.d. samples $w_i^c \sim \text{Gamma}(a_c, b_c)$ and $w_i^u \sim \text{Gamma}(a_u, b_u)$ for $i = 1,2,\dots,N$ and compute samples $R_i^c = \frac{w_i^c}{w_i^c + N_u}$ and $R_i^u = \frac{w_i^u}{w_i^u + N_c}$. We then compute estimators for beta parameters using the method of moments. Specifically, we compute the sample mean $\bar{R}_u = \frac{1}{N}\sum_{i=1}^{N} R_i^u$ and sample variance $\bar{V}_u = \frac{1}{N-1}\sum_{i=1}^{N}(R_i^u - \bar{R}_u)^2$, then approximate beta parameters as $\alpha_u = \bar{R}_u\left(\frac{\bar{R}_u(1-\bar{R}_u)}{\bar{V}_u} - 1\right)$ and $\beta_u = (1-\bar{R}_u)\left(\frac{\bar{R}_u(1-\bar{R}_u)}{\bar{V}_u} - 1\right)$. The beta parameters (α_c, β_c) were estimated analogously.

In the "symmetric" model both the US and China have stakes distributed as $\text{Gamma}(1,4)$. This means when a player prevails in a crisis they are most likely to get no more than 10 "utils" (relative to the cost of the number of adversary nuclear weapons they could employ). With $N_u = 750$, China's baseline resolve is almost certainly below 0.02, meaning they are at most willing to tolerate a 2% chance of nuclear conflict in a crisis. With $N_c = 536$ land based nuclear weapons in 2036, the U.S. baseline resolve looks like this almost certainly below 0.03. For the "asymmetric" game we left the U.S. stakes at $\text{Gamma}(1,4)$ but increased the stakes of China to $\text{Gamma}(1,20)$ to model a much more resolute China in a regional crisis the U.S. may not have the same stakes in.

Below we provide all parameter values used in the tables in the main paper. In all tables U.S. baseline resolve and damage limitation vary across years and are fixed for each level of Chinese damage limitation. The beta distribution parameters (α, β) that characterize the U.S. baseline resolve across years is given in Table 1. The beta distribution parameters that

characterize the U.S. damage limitation across years is given in Table 2. Note that U.S. damage limitation is different under the two different assumptions on Chinese launch under attack (LUA), so we have two different sets of parameters for 2036.

Year	α	β
2010	1.16	13.76
2021	1.06	30.04
2026	1.02	98.18
2036	1.0	134.28

Table 1. U.S. baseline resolve for all tables in main paper.

Year	α	β
2010	56.76	3.71
2021	116.55	10.20
2026	417.62	13.29
2036	655.82	20.04
2036 w/ LUA	323.11	311.6

Table 2. U.S. damage limitation for all tables in main paper.

In the symmetric stakes games the parameters describing Chinese baseline resolve are $\alpha = 1$ and $\beta = 191$. In the asymmetric states games the parameters describing Chinese baseline resolve are $\alpha = 1$ and $\beta = 40$. In all tables the Chinese damage limitation is given in Table 3.

Table 3. U.S. damage limitation for all under no Chinese launch-under-attack.

Mean Damage Limitation	Low Variance DL	High Variance DL
0.1	α: 8.9, β: 80.1	α: 1.0, β: 9.0
0.3	α: 62.7, β: 146.3	α: 1.0, β: 2.33
0.5	α: 124.5, β: 124.5	α: 1.0, β: 1.0
0.7	α: 146.3, β: 62.7	α: 2.33, β: 1.0